DOCUMENT RESUME

ED 340 773 TM 018 029

TITLE Applications in Educational Assessment: Future Technologies.

INSTITUTION Bank Street Coll. of Education, New York, NY. Center for Children and Technology.

SPONS AGENCY Congress of the U.S., Washington, D.C. Office of Technology Assessment.

PUB DATE Feb 90

NOTE 75p.; Contractor report prepared for the Office of Technology Assessment titled "Testing in American Schools: Asking the Right Questions." For related document, see TM 018 025.

PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Academic Achievement; Computer Simulation; Computer Software; *Computer Uses in Education; Educational Assessment; Educational Change; Educational Improvement; *Educational Technology; Elementary Secondary Education; Futures (of Society); *Interactive Video; Models; *Multimedia Instruction; *Student Evaluation; *Technological Advancement; Word Processing

IDENTIFIERS *Intelligent Tutoring Systems

ABSTRACT

The development of improved and alternative methods of educational assessment should take advantage of technologies that enable different aspects of learning, teaching, and student achievement to be part of an improved assessment system. The current understanding of knowledge assessment, new approaches to assessment, and technologies that may enhance assessment are discussed. Emphasis is on computer-based and visual technologies that may contribute to new assessment methods and scenarios concerning the ways technology-enhanced assessment may function in the social and organizational context of schooling. Five categories of technologies that support key features of new assessments are: (1) intelligent tutoring systems; (2) simulations and modeling programs; (3) multimedia systems; (4) tool software (such as word processors and writing environments); and (5) video technology that can record student responses. Three scenarios for the implementation of new technologies in assessment are presented. To the extent that new assessment tasks can sample a broad range of skills, the opportunity for students from different groups to display competence is enhanced. A 72-item list of references is included. An appendix lists 10 examples of educational software from the 5 categories. (SLD)

Applications in Educational Assessment: Future Technologies

Submitted to: Office of Technology Assessment

Center for Children and Technology
Bank Street College
610 West 112th Street
New York, New York 10025

February 1990

This contractor document was prepared for the OTA assessment entitled *Testing in American Schools: Asking the Right Questions*. This is being made available because it contains much useful information beyond that used in the OTA report. However, it is not endorsed by OTA, nor has it been reviewed by the Technology Assessment Board. References to this should cite the contractor, not OTA, as the author.

Introduction

There is now a consensus that the country's educational system is in crisis. Substantial changes are needed throughout its components so that students emerge will from schools with complex and flexible knowledge and thinking skills across subject areas. For example, considerable attention and resources are now being directed to reform of teacher preparation and professional development; curricula, media and materials; parental support; restructuring schools, and so forth. A major and powerful component that needs to change if the system is to improve overall is the way that students are assessed. Considerable attention is now being paid to the reform of testing in this county, going beyond multiple choice testing that emphasizes facts and small procedures, to the development of methods for assessing complex knowledge and performances. A number of testing organizations, states (e.g. California, Connecticut, Vermont), research groups, and schools are now engaged in projects to build and examine new assessment methods. Special emphasis is being placed on performance-based and portfolio approaches to this problem--methods that currently appear to have the most potential to reveal the development of complex/higher-order thinking goals (see below). The interest in new assessment arises for achievement testing of individual students, but also to satisfy accountability requirements for schools, districts, and states.

However, much of the conversation about the development of new assessment procedures is taking place independently of the discussions of technology development and integration in education. Commonly, technology is viewed conservatively as a more efficient delivery mechanism for tests that closely resemble traditional methods. We believe that the development of alternative assessment methods should take advantage of technologies that enable different aspects of learning, teaching, and student achievement to be part of an improved assessment system.

In addition, these new approaches to assessment have important--and relatively unexplored--consequences for the social and organizational aspects of testing and schooling. They commonly require more time and effort on the part of students, teachers, and testing personnel. The methods and testing contexts often piggy-back on--or merge with--the curriculum and products students create in their actual schoolwork. In contrast, our traditional assumptions about assessment are that achievement and accountability tests take place at times and in contexts that are formally marked and separated from everyday schooling.

Considerable research and development is now needed to design and test these methods--with or without technology-enhancement. We believe that current and future technologies may support more valid assessments of complex thinking skills because of the kinds of learning environments they can promote, and because they can record different aspects of performance than are available through paper and pencil means. Technologies may make alternative assessments practical and efficient on a broader scale than would be possible without them.

In this paper, we will take a look at the current situation with respect to current understanding of knowledge development, new approaches to assessment, and the technologies that may enhance new assessment practices. We will take up the following topics in turn:

(1) Assumptions that underlie our approach to new forms of assessment;

(2) Current understanding of learning and knowledge development;

(3) New approaches to assessment;

(4) Computer-based and visual technologies that may contribute to new assessment methods;

(5) Scenarios concerning the ways technology-enhanced assessment may function in the social and organizational context of schooling.

(1) Assumptions about assessment and schooling practice

We begin with a summary of the assumptions that underlie our analysis of the future use of technology in a new assessment systems and schooling practices. There are three components in this overview. First, successful assessment must be systemically valid. Properties of systemic validity--which is being increasingly referred to in discourse about changing assessment--are outlined. Second, key features of needed assessment environments can be synthesized from a range of views about alternative assessment. Finally, the general properties of computer and video technologies that make them valuable media for assessment are outlined.

1.1. Systemically valid assessment

Our ways of assessing student learning are now beginning to undergo the scrutiny required to undertake major redesign. A newly designed system must accurately measure and promote the complex thinking and learning goals that are critical to students' academic success and to their eventual sustainable contributions. The creation of new assessment methods and practices is a systemic problem. An educational system adapts itself to foster the development of the cognitive (and social) traits that its tests or other assessment devices are designed to measure. A systemically valid test is one that induces the curricular and instructional changes required for a system to effectively achieve its valued learning goals (Frederiksen & Collins, 1989; Collins, Hawkins & Frederiksen, 1990; Hawkins, Collins & Frederiksen, 1990).

Assessment is thus a critical component of a dynamic educational system. Properties of systemically valid assessment include:

Directness. In direct tests, the cognitive skill that is of interest is directly evaluated as it is expressed in the performance of some task. Preferably, this is an extended task in which the student uses the skill as part of a motivating and "authentic" performance. For example, the construction of an argument in a legal brief, the design of a building, the development and testing of an hypothesis can be these kinds of tasks.

Scope. The test should cover all the knowledge, skills, and strategies required to do the complex activity, and more broadly, that are valued by the educational system. The instructional system will tend to promote-will adapt itself to promote--those skills that are emphasized by its assessment practices. This has two consequences. If particular types of performances are used for assessment purposes, those modes will be favored in classroom practice. For example, if arithmetic computation is the dominant subject of testing, paper and pencil performances and correctness will be dominant in instruction. If, however, communication of mathematical ideas in language is also a dominant goal, then instruction will be adapted to provide circumstances for practice (the State of Vermont is featuring this aspect of mathematical performance). For any test that emphasizes complex performance, the scoring system needs to take into account the scope of all valued skills. A comprehensive scoring system must include evaluation of all skills required for successful performance.

Reliability. The scoring system must be reliable such that different judges assign the same score to an assessment. Fairness in assessment must be maintained. This criterion has consequences for the definition of tasks used in assessment (e.g. how much can the tasks used to assess complex performances vary across individual performances), and for the nature of the technology and scope of a scoring system (e.g. automatic scoring in an intelligent tutoring system--can all essential aspects of performance be reliably measured?)

Transparency. The assessment system must be clear to those being judged. This includes both teachers and students. Criteria for judging successful performance must be available and understood if the measurement system is to have its most direct systemic effects. The criteria--if they have sufficient scope--can then be clearly used to genuinely improve performance. The terms in which students are judged must be clear to them if the test is to be successful in motivating and directing learning, and in helping teachers to successfully guide practice.

Tests make concrete what is valued by the educational system. The construction of future assessment devices needs to take direct account of their impact on educational practice. People need goals as to what they should be learning. One of the biggest effects of tests and other measurement devices is that they can encapsulate abstract learning goals in a very concrete form--a form that is uniquely motivating to those who are subject to its judgments. We can no longer view our testing apparatus as an unobtrusive means to indirectly monitor students' progress through traditional media--paper and pencil items that largely sample small or inauthentic segments of complex performances. This approach to testing emphasizes the wrong kinds of learning and performance goals.

Thus, not any test will do. "Unobtrusive" written tests of students' problem solving skills administrated in isolation from the curriculum during "testing weeks" are not adequate to the problem. Current views of learning and successful educational practices lead to the specification of particular kinds of instructional activities, tasks, and social organizations in classrooms. For example, these include student engagement in complex tasks that requires thinking and strategic application of skills to a variety of interesting problems, collaborative work among students, increased emphasis on individualized coaching on the part of teachers, and so forth.

In light of accumulated research evidence about how complex learning occurs, monitoring of students' accomplishments emerges as a difficult problem both conceptually (e.g., how do we trace and interpret thinking processes?) and practically (e.g., how do we accommodate time-consuming assessment performances in an already dense curriculum?). The processes involved in thinking, learning and strategic application of knowledge, and the complex products that result are being seen as the new media for adequate assessment. Rather than comparative ranking of students on scales of performance, this view calls for more complex and differential profiles of students' development-in-process.

A new way of thinking about the relationships between instruction and assessment is ultimately needed in schools. This entwining can take different forms (see below), that are just beginning to be worked out. It is not just for practical reasons of time--features of the desired assessment environments include work on complex tasks, and evidence of revisionary effort--but because the entwining goes beyond the notion of simple measurement for monitoring, to use in instructional diagnosis.

We first briefly outline current approaches to learning that support the need for new forms for assessment.

(2) <u>Current understanding of learning and knowledge development</u>

There are many discussions underway about the type of assessment methods that need to be developed to accurately reflect complex thinking/learning/problem-solving goals now at the forefront of educational change initiatives. These discussions generally take place in one of two settings: critiques of current assessment practices (emphasizing the inadequacy of multiple-choice paper-and-pencil technologies); current understanding of learning and effective pedagogical practice. Theories about the nature of mind and the conditions for growth are influencing the discussions of changing measurement practices. While a complete review of these approaches and their potential impact on our testing apparatus is beyond the scope of this paper, we will summarize central ideas and evidence that lead to arguments for alternative assessment methods.

2.1. Research on thinking and learning

Complex thinking and learning can be characterized as nonroutinized – including the development and deployment of skills that allow students to successfully engage in tasks that have multiple solutions, interpretive and nuanced judgments. Students impose meaning and regulate the strategies in nonroutinized processes as they work toward solutions and complex products (e.g. Resnick, 1987a &b). There has been increasing interest in the teaching of such skills, and in adapting educational settings to support their practice (e.g. Baron & Sternberg, 1986; Chipman, Segal & Glaser, 1985; Nickerson, Perkins & Smith, 1985). This is in part because of disappointing evidence about students' abilities in these regards, and in part because of the recognition of the kinds of adaptive abilities that are required by our current and project future workplaces. There is increasing pressure to provide students with the tools and aptitudes that will facilitate lifelong learning rather than narrow--and often inert and inflexible content objectives (Brown, Collins & Duguid, 1989).

There is now a body of persuasive evidence that complex thinking is inseparable from a deep and integrated knowledge base in specific domains studied (especially in science and math: Bashkar & Simon, 1977; Chi et al, 1987; Glaser, 1984; Greeno & Simon, 1986; Tuma & Reif, 1980, but also in social science: Voss, 1986; Spoehr & Katz, 1989). Reasoning and thinking skills and content knowledge are acquired together. Experts and novices differ not only quantitatively in their knowledge and skills, but qualitatively in the ways their knowledge is organized and the kinds of skills and procedures they use to reason with. Central to the acquisition of complex cognitive skills and comprehensive domain knowledge are practices that help students to use, and thus make usable, their knowledge. Key research on reading (Beck & Carpenter, 1986; Brown & Day, 1983), writing (Bereiter & Scardamalia, 1986; Collins, Gentner & Rubin, 1981), mathematics (Schoenfeld, 1985) and programming (Anderson & Reiser, 1985) has demonstrated the importance of knowledge using practices for text comprehension, recognition of problems, integration and application of information, managing parts of a complex task, representing and revising problem subcomponents and solutions. In addition, knowledge is actively constructed and used in complex circumstances. Current research is investigating a view of learning and expert performance as situated in circumstances that include webs of distributed knowledge. Cognitive performances in real world settings are inextricably supported by other people and knowledge-extending artifacts (e.g. computers, calculators, texts and so forth).

This point of view challenges traditional views of how to determine students' competence. If knowledge is complexly tied to situations of use and communities of knowers, then compiled lists or matrices of abstracted concepts, facts, procedures, or ideas are not adequate descriptors of competence (Resnick, personal communication). Achievement needs to be determined by performances or products that interpret, apply, make use of knowledge in situations. We cannot be assured of learners' competence by their responses to decontextualized items or abstract problems. This requires new circumstances for assessment--circumstances that we do not yet know how to do an a large scale.

Today, students often learn individual facts about people, places or events without understanding how they are integrated or why they are significant. In the absence of such understanding, students' knowledge will be unavailable in situations that are different from those in which it was initially acquired (e.g. Pea, 1987). Students have great difficulty applying knowledge they can only remember, rather than working with it in a way that supports integrative problem-solving (Pea & Soloway, 1987). Young learners tend to acquire knowledge as discrete detail and need to be supported with strategies that develop interconnected, coherent working knowledge (Scardamalia & Bereiter, 1984). Instruction focused primarily on textbooks and multiple-choice tests tend to encourage poor kinds of knowledge representation (Scardamalia & Bereiter, 1986). Knowledge is best developed in tasks for which it is functional.

Tests are one of the great incentives that motivate study. But they often lead students to nonproductive study strategies. In learning information for today's tests, students are often developing primarily rote strategies and memorizing information that is inert and is of little use for non-school problem solving. Memory strategies can be useful, but when information and procedures are learned in isolation, apart from the contexts in which they might be used, students cannot transfer what they have learned to new relevant problems.

This accumulating evidence supports four characteristics of instruction that can contribute most productively to student learning. Instruction should be individualized, adaptive, interactive, and located in meaningful tasks.

(3) New approaches to assessment

Current views of expertise (the nature of knowledge representation and strategic use in domains as complexly integrated) and learning (constructive, situated and transformatory of existing representations involving generative and varying practice on relevant problems) have direct consequences for a new generation of assessment. Historically, test development has been at least in part separated from these kinds of research currents. As noted by Haney and Madaus (1989),

> Theories of cognition do not currently play a central role in test development. As a result, the task of explaining even so long-recognized an ability as reading comprehension in terms of a theory of information processing or other advanced psychometric concepts has barely begun. If more psychologists who worked on theories of cognition also worked on test development, testing today might be further advanced. This kind of research is difficult, but the effort is needed (Committee on Ability Testing, National Research Council, 1982).

The concept of systemically valid testing is based on a much closer merger between theories of learning, instruction, and the design of learning environments, and the design and practice of assessment. To illustrate current views about alternate assessment in relation to educational practice and cognitive theory, four representative voices will be summarized. They share the central feature that assessment must elicit and directly measure complex accomplishments. Each approach explicitly or implicitly embodies commitments about the nature of intelligence and instruction.

In creating or adapting new measures--in our case, those that take advantage of technologies--these underlying commitments need to be taken into account. They have an impact on the nature of the tasks that students are asked to do, the kinds of records that are collected about their performances, and the design and features of the scoring systems that guide and judge their work.

The first two perspectives are oriented toward performance-based and portfolio assessment, which require particular kinds of records of student work. As noted above, the development of these methods only rarely is taking place in concert with efforts to use technologies for problems in education. In general, without regard to technologies, the work in alternative assessment is most advanced for writing assessment, and for the creative arts.

We believe that these new forms of assessment can well take advantage of technologies to structure assessment tasks, and collect performance records and products in many subject areas. The latter two perspectives presented here go beyond task and record structuring, to capturing real-time detail of students' actions and interactions (intelligent tutoring approaches, and video technologies).

3.1. The nature of intelligence and portfolio assessment. Wolf, Bixby, Glenn, Davidson & Gardner (1990) argue for a changed understanding of the nature of intelligence than that which underlies traditional approaches to assessment. They provide an analysis about how our cultural understanding of mind and individual difference has conditioned a view of testing for intelligence as a fixed, isolated, relatively unchangeable capacity. Tests are thus constructed to be comparative performances of students on small decontexted items.

Their view of testing is based on a notion of intelligence as fluid, adaptable, and applied, with differential profiles characterizing different students. We therefore need to transform our system from a testing culture--which has as its goal simple and comparative measurement of educational progress by ranking and classifying on continua--to a culture of assessment which is based on complex documentation of progress through profiles of growth. Adequate assessments are based on students' accomplishments, and the conditions under which these accomplishments are possible, rather than limited and decontexted responses to small fixed items.

This approach to assessment relies heavily on methods that involve collection of bodies of products, and works-in-process to demonstrate progress over time, repeated and revisionary attempts at the same tasks or products. Assessment procedures are also needed that accommodate the nature of real world performances to include social support, and use of tools and other artifacts Scoring procedures are based on clinical judgments. Student progress should be considered in terms of passing milestones rather than compiled correctness of response to batteries of items. "We can deliberately honor real world knowledge; expressly scaffold the relevant steps or questions in the process; permit students the stimulation of collaborative work; allow for a range of forms and formats in responses (e.g. graphs, essays, etc.),. and provide for revision." Two examples of alternative assessment methods illustrate this approach. The first is a kind of performance-based assessment that permeates the instructional context.

The national assessment in Holland structures performance-based assessments for students by designing problems for the year-end examinations. For instance the committee of teachers in art history selects a unifying subject (like "revolution"). Students are provided with information packages to guide their study of art throughout the year in ways that help them to critically develop the theme (e.g. readings, lists of museums and so forth). Teachers are encouraged to work with students in a variety of ways (discussing, modelling processes, helping them to develop individual interpretations and points of view). This assessment approach supports students in doing individualized in-depth synthetic work, yet around a context of shared ideas, procedures and a common generative problem.

The research program of Wolf et al has adapted and developed portfolio-based assessment, and created the notion of process-folios that goes beyond the collection of finished works to collection of records of works in progress. These measurement systems move away from a notion of pure and decomposed measurement of skills to a model of clinical judgement applied to complex and individualized productions, and can include productions in multiple modes or media. Portfolios are methods for keeping records of students' ideas, work, and revisionary activity as they make complex creative products. Records of this sort are particularly rich sources of information about progress in individualized and difficult to measure projects or creative endeavors. Portfolios rely on human judgments about products of students' problem-solving. An important part of the assessment information can come from student judgments--how they make selections for inclusion of their work in portfolios and their processes of reflection as they look over the work. The development of such techniques for assessment practices can draw on features of technologies for aspects of task definition, record keeping, and scoring.

Performance and portfolio-based assessment projects of this sort are underway in Connecticut (science), California (writing and mathematics), Vermont (writing and mathematics), New York (science), England, and Pittsburgh public schools.

3.2. Authentic assessment in schools. Wiggins (1989) has been a key spokesperson in the educational community, and has consistently articulated the properties of authentic assessments for schools. Authentic assessments illuminate the capacities of a wide range of students, and, like Wolf et al, are based on students' complex accomplishments. They need to do more than measure whether learning has occurred, but model sustained thoughtfulness and offer useful diagnostic feedback to students and teachers. A valid test in this view is not retrospective about learning that has occurred. The test itself is educational in the sense of improving performance. Testing and the criteria to which people are held accountable should be viewed as raising standards of performance--and thus motivation to improve--rather than inert judgments about ability.

In this view, authentic assessments have the following properties (Wiggins, 1989). Structure and logistics. They are: public; involve multiple criteria for success; can involve collaborative work; recur. Valid assessments also involve repetition. One test result cannot signify mastery or its absence. A profile of practice and revisionary attempts is required. They are significant tasks that are worth practicing; do not involve arbitrary time constraints; give students clear and detailed feedback about their performances. Intellectual design features: Assessments are contextualized and present complex challenges; stress depth of accomplishment; involve ill-structured and ambiguous tasks; may provide scaffolding to see what students can accomplish with support; require engaged use of knowledge from a wide repertoire; are representative challenges within a discipline. Scoring and grading. Criteria for success are complex and require judgment rather than simple counting of errors. Primary trait scoring may be appropriate. Tests are scored with reference to authentic standards in the discipline, and are designed to enable students to demonstrate what they can do. Scoring results in multifaceted profiles rather than single aggregates. Fairness and equity. There is room for differential style and ferreting out of students' strengths; comparisons are minimized; students are "scaffolded up" to successful performance.

This perspective also leads to assessment forms that are based on complex task performance. Portfolios and performance-based assessments may meet these requirements. In the most extended case, student assessments are based on "exhibitions"--long term comprehensive projects that are a core feature of the curriculum and through which students are assessed not only on their finished products, but on the explanations that they provide when queried by a teacher or panel about their work (cf. Coalition of Essential Schools, Wiggins, ms). Alverno College has also developed procedures and criteria for this approach to assessment.

3.3. Cognitive science and assessment. Another kind of discussion of the requirements and forms of alternative assessment grows directly out of cognitive science research (largely in the areas of science, mathematics and computer science). Frederiksen and White(1988), for example, argue that the quantitative view of mental measurement is challenged by the accumulated evidence from problem-solving studies in a variety of domains. They suggest a view of expertise and learning organized by representations of mental models. Learning occurs through transformations of inadequate mental models to canonical representations of knowledge, and using this knowledge in different contexts. The goal of measurement is not a quantitative scale (it obscures the underlying models), but representations of students' mental models for reasoning and their strategies for applying knowledge to problems. The metaphor for measurement is a "space" of mental models and the conditions of their transformation rather than a two-dimensional scale. Expertise is represented as a set of linked alternative models (qualitative, quantitative, functional and so forth). Growth is viewed as an evolutionary progression of mental models. Instruction takes the form of extended problem-solving episodes, where feedback and explanation is provided to students to induce model transformations. The approach supports multiple strategies in problem-solving, and can be used to assess learning potential (how much support is needed for model transformation) as well as current state.

The approach to assessment that follows from this view leads to the development of techniques that reveal the current model state of the student, and the conditions of her further development. Records that reveal sequences of actions taken by students, reasons for taking those actions, learning strategy preferences and learning times are the basic data. Extended problem-solving scenarios need to be captured, and used as the basis for understanding students' knowledge state and potential for transformation. Students are judged in through comparisons to explicit models of expertise.

The goal of assessment is to place students in generative rather than evaluative situations, and to understand their actions in terms of models of expertise. The result is a profile of students' competence.

This approach to learning and assessment is most efficiently embodied in intelligent systems, since explicit models of expertise can guide students' actions in a problem-solving environment, and sequences of actions can be automatically accumulated as the basis for analyzing knowledge state. Assessment information is based not on single responses, but aggregations of responses (sequences of actions over time), from which current and potential knowledge states can be inferred.

3.4. Diagnostic assessment. A program of research has been underway for over a decade concerning the use of diagnostic instruction in classrooms (its occasions and the kinds of often tacit information that teachers seek as the basis for instructional action). This notion of dynamic assessment focuses on the information that is yielded by identifying the next instructional steps needed to move a child from a current display of competence to the next level. Considerable research has demonstrated that this kind of careful, individualized instructional scaffolding is both very supportive of learning, and a powerful source of understanding the level of students' competences in complex tasks. "The notion of supportive learning environments to reveal and develop a child's potential to its fullest extent cannot help but influence how teachers assess a child's competence and structure instruction. What might give teachers pause are practical problems in implementing dynamic procedures on a wide scale. (Brown & Campione, 1986; Campione & Brown, 1985). Dynamic assessments that make use of social scaffolding by more knowledgeable individuals can be used to accurately represent current knowledge states as well as directly guide instruction.

3.5 Summary. These programs of work in developing ideas and techniques for alternative assessment are grounded in different locations and different purposes in the research and educational communities. They have somewhat different implications for assessment procedures. But there is considerable overlap in the key qualities of the assessment conditions that are proposed. Taken together, they further articulate the properties of systemically valid tests in relation to theoretical and practical evidence about the development of thinking and problem-solving.

Key Qualities of New Assessments

(A) Directness

1. Assessment must be based on performance of complex tasks, with sequences of responses or complex products as the data.

2. Performances occur over extended time periods or several sessions, appropriate to the task requirements and individual differences (eliminate arbitrary time restrictions).

3. Tasks can be repeated, and include revisionary activity.

4. Assessments take into account the situated nature of expertise in terms of use of social resources (collaborative learning) and cultural artifacts (performance-supporting tools).

(B) Scope

5. Tasks support multiple approaches and solution strategies.

6. Assessments include learning potential as well as knowledge state (notions of scaffolding and diagnosis).

7. Assessments include not single tasks, but multiple views of students' capacities.

8. Assessment should include social/interactive skills concerning knowledge use and communication in addition to the traditionally emphasized cognitive skills.

(C) Transparency

9. Feedback about performance is given to students throughout the task, or afterwards, with the goal of improving performance.

10. Feedback should be usable for instructional purposes.

(D) Reliability

11. Scoring/judgments are made in relation to models or criteria for expertise or high quality performance rather than relative ranking among students.

12. Scoring or judging procedures result in profiles rather than single comparative scores--they may include clinical approaches, primary trait criteria, benchmarks, explicit models of expertise in a domain.

Many of these features are difficult or impossible to achieve in schools with traditional instructional and assessment media. Technologies offer the means to realize many of these features as part of future assessment systems.

(4) Properties of computer and visual technologies as media for alternative assessment

Electronic visual and computer technologies – and systems that electronically combine the them – provide opportunities for creating assessment environments and recording new aspects of students' performances. They make it possible to construct assessments that more fairly represent the range of knowledge and skills toward which education should be directed.

These technologies also provide the means to operationalize many of the key features of measurement required by current views of learning and mind, and assessment reform. An assessment system that is faithful to the broad scope of learning outcomes known to be critical must assess other processes and performances than those that can be measured by paper and pencil tests. For example, some critical skills and abilities that cannot be so measured include how well students listen and ask questions, how they handle new facts in their theories and questions about consistency, how well they formulate and test hypotheses, how well they make oral presentations and arguments that make use of visual and graphic media as well as verbal and so forth. Technologies may also be able to support these general features of new assessments:

--diagnosis of individualized and adaptive learning;

--repeated practice and performance on complex tasks and on varying problems with feedback;

--recording and scoring multiple aspects of competence;

--maintenance of an efficient, detailed and readily updatable "history" of performances.

Technologies add an important new capacity to an assessment system for evaluating these aspects of students' performances. They make new "slices" of performance available to the lens of assessment. New ways of eliciting, recording and judging thinking performances need to be developed. The research and development process should be designed in such a way it takes full advantage of the possibilities offered by the various technologies, rather than considering them after-the-fact.

4.1. General properties

We will first describe the general properties of computer and video technologies that make them attractive components of new assessment systems. The particular types of software and systems within these categories that are especially promising will then be discussed and briefly exemplified. More detailed specification of the examples is available in the Appendix.

4.1.1. Computer technologies contribute important new capacities that may be adapted to advantage by a new assessment system. These new features can be described in three ways.

Thinking processes. First, computers make available thinking processes as a students' thinking and work evolves over time. They enable certain kinds of process records to be kept about students' work on complex tasks as it evolves and is revised in single or multiple work episodes. Thus, in relation to new key features of assessment practices, they may allow the efficient capturing of views of students' performances in problem solving that are otherwise evanescent, cumbersome to record, or invisible. Because they can trace process, computers can provide the structure for tasks and record the processes by which students maneuver through a problem (Collins, 1990a; Frederiksen & White, 1990). For example, it is possible to keep records of whether students systematically control variables when testing a hypothesis, to look at their metacognitive strategies (Collins & Brown, 1988; Schoenfeld, 1985), to determine what they do when they are stuck, how long they pursue dead ends, and so forth. The ability to trace the problem solving process allows computers to capture the strategic aspects of students' knowledge.

Learning with immediate feedback. Because it is possible to put students into novel learning environments where the feedback is systematically controlled by the computer, it is possible to assess how well or how fast different students learn in such environments, how they use feedback, how efficiently they revise (Collins, 1990a). This offers the attractive property of practically embodying dynamic assessment techniques in an assessment environment. Thus, this dynamic scaffolding offers a way of going beyond simple measurement of performance level, to a more sensitive measure of learning potential and strategy in a particular domain.

Structuring and constraining complex tasks. Computer environments can be used to structure--and constrain--students' work on complex tasks in ways that are difficult to otherwise achieve. In the case of simulations, dynamic problems that may have multiple outcomes can be can be carefully designed, and students' progress toward solution can be automatically recorded (including time, strategy, use of resources and the like). The simulated tasks may be modified to change their complexity, or repeatedly attempted by students as a way of recording progress over time. The tasks can be designed to record students' abilities to deal with realistic situations, like running a bank, repairing broken equipment (Collins, 1990b), solving practical problems that embed mathematics (e.g., *Jasper*, below, Cognition & Technology Group, Vanderbilt University). They can provide students with complex information sources that can be consulted as they work, creating data about not just what students know, but about how they sift, interpret, and apply information. It is thus possible to measure students' abilities in understanding situations, integrating information from different sources, and reacting appropriately in real time. Paper and pencil, or video, cannot really simulate situated problem-solving, so with respect to measurement problems, only computers and the computational properties of multimedia systems give us a view of people's practical intelligence as they apply it in real time.

Using models of expertise. In the case of systems with some degree of intelligence, models of expertise can be used to guide and gauge students' development of domain knowledge and strategy. In this case, learning and its monitoring are simultaneously occuring through the intervention of an expert system that diagnoses students' level of competence in a complex space of expertise. This allows not just a recording of the problem-solving process, but the juxtaposition of students' process with a detailed model of strategic expertise in a carefully analyzed domain.

Thus, the computer features particular strengths (e.g. structuring complex tasks and tracking the process of development and thinking) that offer several ways to tap students' abilities that are inaccessible with other media.

4.1.2. Video technology allows us to look at the interactive capabilities of people, and the way in which they may use aspects of their social and physical environment in accomplishing tasks. Up to now, individual and isolated cognitive performances have been the sole determinant of competence. With recognition of the basic social nature of learning and work throughout our society, these abilities must be brought into central significance. Video technologies can--now that production as well as consumption capabilities are broadly and economically available--record ongoing activities, products at various stages of development, explanations, interactions and presentations in rich detail.

4.2. Categories of technologies for alternative assessment

In this section, we will describe more specifically five categories of computer and visual technologies that support key features of new assessment practices. We will first describe the properties of the general category as it relates to assessment features, and then briefly illustrate with examples of existing commercial or prototype software. More complete descriptions of specific examples are found in the Appendix.

It is important to note that it is relatively rare that a technology-based system has been/is under development for solely assessment purposes. With the exception of some technology-based systems to deliver relatively traditional forms of testing, most work in this area is attempting to make use of available software (commercial or research prototypes), retrofitting or adapting it to assessment purposes. One form of adaptation, for example, is to modify simulation or tool-type programs to automatically collect and store records of students' work with the software.

4.2.1. Intelligent tutoring systems (ITS). Intelligent tutoring systems provide intelligent, individualized guidance to students as they learn through working on problems. These systems grow out of the considerable research and development that has been devoted to artificial intelligence and the development of expert systems in the last decades. Intelligent tutors take advantage of this work by combining models of domain expertise, with models of learner(s) and pedagogical techniques. They are designed to diagnosis a student's learning level or operative mental model by comparing her performance on tasks with some representation of expert performance. Appropriate guidance is then offered to the student. Different systems are based on different types of models of expertise (see Appendix). They provide different kinds of pedagogy and guidance to students. For example, some select problems appropriate to the student's level, and offer process hints and queries. Others diagnose errors in students' work and identify and explain them.

An ideal ITS system would be capable of autonomous pedagogical reasoning--that is, combining expert knowledge with explicit understanding of the development of knowledge in a domain, and appropriate pedagogical techniques. Ideal systems have yet to be produced, but several interesting prototypes are being used as research environments in instruction and assessment. In these intelligent tutoring systems--unlike computer-assisted instruction--knowledge itself is explicitly represented and used to guide dynamic instructional interactions that are unique and sensitive to the level of the learner.

> For research on instructional systems involving artificial intelligence, the purpose is not to provide a software-engineering framework within which experts can compose instructional sequences by representing their decisions in the form of programs *[i.e. computer assisted instruction]*. Rather, the purpose is to capture the very knowledge that allows experts to compose an instructional interaction in the first place. Instead of decisions resulting from some knowledge, it is the knowledge itself that is explicitly represented so that it can be used in computer-based systems. It is then the responsibility of programs to compose instructional interactions dynamically, making decisions with reference to the knowledge with which they have been provided. (Etienne Wenger, 1987)

These systems require considerably more powerful and more expensive hardware than is commonly available in schools today. It is unlikely that most schools are likely to acquire this advanced computing power in sufficient quantity foreseeable future, even if hardware costs decrease significantly. The small number and research status of most systems means that their use is not now a factor to motivate hardware purchase.

Assessment and ITS. ITS were originally conceived as instructional systems and only recently are they being explicitly adapted as assessment environments (see below for two examples of ITS that have been partially adapted as intelligent testing environments--PROUST and QUEST, Littman & Soloway, in press). Their designed capacities for detailed diagnosis and intervention through analysis of aggregated sequences of actions make them especially attractive in this regard.

Intelligent tutoring systems can closely monitor and measure student progress in the context of instruction. Such assessments can be directly based on problem-solving tests, where problems of differing difficulty are given to students and their solution steps closely monitored in relation to a model of expert performance. Each system is of course based on a different model of expertise in a domain, but more interestingly, each also adopts a different approach to the nature of representation of expertise, and thus the kinds of traces of performance and model of development that is available to the lens of assessment (see below).

Such systems make it possible to assess new components of learning and problem solving. They can structure problems in areas such as models of gas diffusions, heat flow, and electrical circuit behavior. As students interact with the systems they carry out extended problem solving scenarios that can be qualitative in nature. They can for example design an experiment to test Boyle's Law, or determine the components of force on a pendulum bob at a particular point, or ascertain the behavior of a light bulb in an electrical circuit as changes in conductivity of circuit components are introduced. Problems such as these invoke not only the <u>use or display</u> of knowledge, but also <u>strategies of inquiry</u> that can create new knowledge. These systems can promote the development of a new form of learning diagnosis, including understanding the form of laws and theories; use of cognitive models in a domain in problem-solving; control and use of problem-solving strategies; processes for generating knowledge.

In ITS, responses of students throughout the learning process can be aggregated and interpreted in relation to representations of expert problem-solving. The systems offer the opportunity to understand student performance not simply in terms of correct answers, but in sequences of responses that can reveal how students learn. Different system designs, however, permit aggregation of different kinds of data about performance--the information useful for assessment purposes. For example, PROUST (a system for tutoring programming in Pascal--see Appendix) identifies students' errors as they generate programming code to meet particular task specifications. QUEST (an ITS for learning about electricity and electrical circuits) records steps in solving a problem, and compares them to a progression of models of expertise. The system can also record how students use explanations, hints, or examination of models as they learn.

Intelligent tutoring systems are not large in number, and are concentrated in three domains (science, mathematics, and computer science) where computational power has the most leverage. It is not clear how feasible such approaches are to intelligent tutoring in other domains of expertise that tend to be more ill-structured like history or literature (e.g. Spoehr & Katz, 1989). Existing ITS are however powerful design examples that can be examined for potential contributions to student assessments. Examples of how these systems might become part of assessment in schools are available below, Scenario 2.

In general, tests would start with easier problems, and as with adaptive testing, depending on how well the student does, subsequent problems would be easier or more difficult, or different kinds of information would be provided to students. As they solve problems, students receive feedback about their performance in accordance with the constraints and representations of the underlying models of expertise and learning, and intelligent coaching toward greater success. That is, the full capability of the tutoring system to teach students can be employed as part of the testing procedure. The test would then measure not simply their prior ability to perform tasks but how well they can learn to perform tasks given appropriate feedback and advice (dynamic assessment).

These systems also have another interesting feature for the purposes of developing new assessment practices. Their development has forced the emergence of new forms of written notation for the recording of the problem-solving process (Wenger, 1987). While trace data and records of time-on-task provide certain types of information about the steps that students take in problem-solving, working with some of these systems has created new ways of representing process information that students must manipulate. For example, in *Algebraland*, students manipulate mathematical operators that then automatically perform the desired calculations; students have constantly available written traces or tree structures of their problem-solving process. These novel notational systems and representations that students manipulate have interesting consequences for emphasis on process--and its capture for measurement--in instruction.

Advantages. Thus, ITS can make different aspects of problem-solving and learning processes available. In the course of doing individualized diagnosis for learning, these systems can accumulate records that automatically supply a picture of student progress on the tasks that are posed. Scoring is done in relation to the model of expertise that underlies the system, and is a detailed and valid assessment to the degree the model accurately characterizes knowledge in that domain.

Limits. There is interesting potential for intelligent tutoring systems to function as intelligent testing environments in learning settings. What are the limitations and risks involved in their successful use as a substantial part of new technology-enhanced approaches to assessment? There are several.

First, as noted above, there are very few systems currently available and they cover quite circumscribed parts of the curriculum. Expanding their coverage would be quite costly, and it is unclear that it would even be possible for significant segments of schooling. To make this broadly possible in the foreseeable future would require significant research and development financing and effort. These systems are extremely difficult to build, and experience with them over the last ten years indicates that the prototype systems commonly go through major design revisions. Few have left the laboratory for the practical world (*Geometry Tutor*, for one, is now being used in some classrooms). A great deal of investment (financial, and the efforts of teams of people over a long period of time) is needed to bring a single system to completion. It is unlikely such monies as will be required to build these systems for the range of domains and developmental levels that must be assessed will be available for this purpose.

This does not mean that they don't have potential as components of new assessment practices. But it is unlikely that ITS will play a major role in testing in schools in the foreseeable future. They may be used as central resources in some courses or for some testing (e.g. the success of the *Geometry Tutor*), but they are unlikely to have a big impact on realizing alternative assessment on a broad or comprehensive scale. In light of potential development costs of new tutors, costs for adapting current systems as intelligent testers, and potential limitations on the domains for which ITS are appropriate, current systems can best be considered as design examples. They may be severely limited with respect to the scope criterion for systemically valid tests.

Second, the available systems were conceived primarily as instructional environments rather than testing environments. While interesting and complex records of students' work are available, the procedures for collecting, handling, compiling and interpreting the most attractive of this material for assessment purposes remain to be worked out.

Initial efforts in this regard and research studies about how to best structure tasks and automatically score complex performances are sobering. As with all assessment procedures regardless of promise, under practical testing and revision, the measurement systems will gravitate toward promoting the skills that they measure well. With ITS, these tend to be algorithmic skills, which may be the less important skills to learn. Research suggests that the tasks that are structured in ITS for assessment purposes may need to be much more constrained to be reliably scored than at first appears--making these technologies much less interesting with respect to directness and scope criteria.

Two examples will illustrate these effects. Research with PROUST discovered that automatic scoring was much more reliable with more constrained and less interesting pr blems. At first, students wrote programs according to specifications. But it was found that these generative performances were much less reliably scored by the system than more limited tasks in which students identified errors in others' work. The tasks may thus be required to be constrained in ways that do not promote generative performance. The system must also work within the constraints of a quite limited number of well-understood problems for assessment.

Another intelligent testing system is under development at ETS for the assessment of architectural design skills as part of a professional licensing process. Because of the limits of automatic scoring, it appears that the kinds of shapes and objects students must work with in their designs, and the actions they can take on them, must be quite restrictive (Braun, personal communication). The problems thus posed are now limited, and feature assessment of basic knowledge and skills, not generative and creative design solutions to architectural problems. This could have the unintended effect of reshaping course content in the direction of the constrained problems that the testing system can automatically score.

Thus, the kinds of performances that ITS may promote in assessment circumstances could have the effect of doing what our measurement system now does. Instead of expanding outward to include complex, generative thinking and learning performances and products, ITS could have the effect of greatly limiting the region of assessment to a very few well-understood problems, and the more tractable but less interesting problem-solving skills.

In an analogy to ice-skating competitions, assessments might be thought about as covering "school figures" and "free skating". ITS may be best suited to careful tracking of "school figures", but there is a resource and effort tradeoff in developing intelligent testers for these purposes. It is likely that they fit limited but important niches in the assessment/instructional environment, but unlikely that they will cover much of its scope.

4.2.2. Simulations and modelling programs

These kinds of software programs are designed to simulate or model a phenomenon (e.g. *Dynaturtle*, which models Aristotelian motion under user-control), or to enable people to engage in simulated problem-solving or decision-making activity (e.g. *The Other Side,* which engages users in geo-political decision-making). Examples of this type of software can be found in most domains, both in and out of school (e.g. STEAMER, which simulates decision-making to control an energy-generating facility, and is used to train workers in difficult plant management with no real-world consequences).

These programs enable people to observe, control and make decisions about scientific phenomenon that would otherwise be difficult or impossible to observe. For example, with *Physics Explorer*, students can specify variables that control various physical motion phenomena (e.g. friction), and then observe the effects of these conditions on simulated motion of objects.

Programs of this sort also enable students to carry out complex sequences of actions by simulating decision-making activity in the sciences and social sciences, history, literature. For example, *Rescue Mission* is a simulation that allows elementary school students to navigate a ship to rescue a whale trapped in a net by learning the mathematics and science required to read charts, plot courses, and control navigation instruments. This kind of software is often popular with and motivating to students (e.g. *Where in the World is Carmen Sandiego* – geography concepts in elementary school), and is found today in many classrooms. *SimCity* enables students to simulate the decision-making involved in planning and building a city, including such features of the infrastructure as electricity, environmental quality, and so forth.

Software in this category is available for almost all computer hardware now available--from the AppleII level of machines that represent a large portion of the presently installed school base, to the most advanced machines used in military training. The sophistication of the program is of course commonly a function of the sophistication of the hardware, but many clever programs that could well be adapted for assessment purposes run on relatively low-level machines.

Simulations can support the study of curricular domains much as real laboratories do in science, or as problems in multiple perspective taking and information integration do in history or literature. They can model and reveal process that would otherwise be inaccessible to experience (e.g. *ThinkerTools*, White & Horowitz, 1987), allowing students to experiment with hypotheses in ways that allow them to simulate real-world effects (e.g. *Dynaturtle*), and to concretely think through, represent, and try out variables that affect phenomena or systems (e.g. *Physics Explorer*). Time can be compressed or expanded so a process that is too slow or too fast to be observed can be studied in detail. Spatial scale can be changed to bring the very large or the very small into comfortable viewing range. Experiments that are not feasible in the real world can be simulated (Tinker, 1987). Systems can be designed and the effects of their operation explored. Students can engage in learning by doing to try out alternatives and engage in "what if..." thinking (diSessa, 1982; Frederiksen & White, 1990a). Students can engage in real, functional, complex tasks, trying out alternatives and reviewing choices, actions, and their consequences.

Dynamic modelling of natural process in artificial worlds can call students' attention to specific features and aspects of a complex phenomenon (e.g. ecosystem dynamics), and can structure their tasks with these complex phenomenon in ways that reveal their problem-solving processes and understanding. Structures can be seen and manipulated from perspectives not possible in the real world. They permit one to begin with highly simplified abstractions, that can become increasingly complex as the students' knowledge progresses. They offer the possibility--in both instruction and assessment--of "achieving an appropriate balance between giving students the direction they need and permitting them the freedom to explore and to make their own decisions" (Feurzeig, 1986).

Assessment and simulations. This type of software has very interesting potential for assessment. Simulations and modelling programs support problem-solving tests/tasks, where, for example, students can carry out complex sequences of actions, and can automatically collect records of their work. Simulations call for the student to generate entire sequences of actions that lead to solution, supporting the exploration of multiple paths, the use of different kinds of information, and the use of feedback on subsequent solution attempts. While the responses allowed are not unlimited--there is usually a restricted class of inputs that a system can process--neither are they single-correct-item, multiple choice response formats. The constraint on response also allows particular structures to be built into the assessment tasks. This has an impact on the kinds of data about thinking process that can be recorded and the nature of scoring systems that can be applied. The responses required by simulations are generative--within bounds--but at the same time can be precise enough to be evaluated according to well-defined criteria.

The basic approach can be applied in any domain where simulations can be part of the instructional system. For example, one computer-based system for teaching geography and history is *Geography Search* (Tom Snyder, Snyder & Palmer, 1986). It is a historical simulation for the Encounter--time after Columbus arrived in the New World. In the simulation, students have to purchase supplies for their trip and navigate to the New World using compass and sextant. They must plan and revise their voyage as they go. Historical simulations like this give students an understanding of the reasons why events take place in a historical context. They can also provide a constrained and structured problem context where students' decision-making processes can be traced as part of understanding their progress.

Simulations may be useful in several ways as parts of new assessment practices. For example, the following types of tasks are of interest:

Problem-solving. Performance-based assessments currently under development are often organized around problem-solving activities where students are required to carry out a complex task and are scored on solutions, or evidence of steps toward a solution (cf. Connecticut Common Core of Learning in science). Simulations, and the emerging multimedia systems (see below), can have key roles to play in structuring tasks and in recording aspects of students' performance otherwise inaccessible to observers.

Simulations (or multimedia systems) can support performance-based assessment tasks that involve understanding the process of students' thinking over time. For example, these are some of the tasks we have considered or tested:

(1) Formulating the relationship between variables. In *Physics Explorer* students can set various parameters associated with a large variety of physics models (one and two body motion, waves, electrostatics and so forth). They can observe the effects of their manipulations on the behavior of these models, plot the parameters and their effects graphically, and write verbal explanations in a note-taking space. One problem might be to figure out what variable affects the period of motion or length of the arc in conducting pendulum experiments. Another problem might be to figure out how variables affect the friction acting on a body moving through a liquid.

In addition to their solutions to problems posed, students' thinking processes can be assessed through the sequences and compositions of variables they test as they try to discover the laws governing their behavior recorded as trace data about their solution steps. The kinds of written records they keep and the graphs they choose to generate can also be part of the evidence that is used to score their performances. These data can be readily gathered from students' problem solving attempts, and allow for multiple views of the learning process in problem solving.

Records of students work might be evaluated in terms of the following traits: how systematically they consider each possible independent variable; whether they systematically control variables while they test a hypothesis; whether they can formulate qualitative relationships between the independent variables and the dependent variables; whether they can formulate quantitative relationships among the variables.

(2) Troubleshooting or diagnosing problems. Another kind of complex problem-solving task that can be structured through simulations and not

readily otherwise involves diagnosing why a system is not behaving as expected. Such problems are most common in computer science, medicine or other applied sciences but they can occur in any system (like government, business, education). Students can be given faulty systems through simulations--such as circuit behavior--and be required to troubleshoot why it is not doing what it is supposed to do. Their performances can be evaluated in terms of (1) how systematically they collect data; (2) the consistency of their hypotheses about the data they collect; (3) how well they test hypotheses to rule out other hypotheses.

(3) Integrating multiple types of information in decision-making. Many problems require the selection, interpretation and synthesis of complex information from different perspective and disciplines in their solution. These kinds of complex decision making tasks are multi-step and often require data in different media and formats. Simulations can readily structure assessment tasks of this sort--tasks that are open-ended, yet constrained by the problem and the data used in its solution. For example, *The Other Side* requires students (in collaborative and competitive arrangements) to consult and apply information about natural resources, economics and international politics to solve a problem between two nations. Likewise, the popular simulation, *Where in the World is Carmen Sandiego?* requires students to apply historical and geographical information to solve a series of problems. Students might be evaluated on (1) how systematically they screen and collect information; (2) how well they reason and apply it to the specific problem at hand; (3) how well they plan; (4) how well and critically they interpret information in different formats; (5) their skills in synthesizing information from different sources; (6) specific domain knowledge.

(4) Learning with feedback. With many simulations, it is possible to give students feedback on what they have done and hints about good strategies to use. For example, in *SimCity* the software simulated the responses of citizens to aspects of the urban design being created by the students. Students are required to respond to this feedback, or fail at the tasks. To emphasize the importance of revisionary activity and improvement as a feature of assessment environments, students might be evaluated in terms of how much their performance improves in some fixed period; how responsive they are to suggestions they receive and how well they use this information; their overall performance level following revision.

(5) Design activities. In addition to supporting the production of student products--expanding the opportunity to use process and revisionary records as part of the assessment--some simulation software can be used to support a type of composition activity, that of design (tools can also be used in this regard--especially graphics and modelling programs). Computers provide a medium in which students can carry out design tasks, like designing a circuit, an ecosystem, or a governmental policy. The designed system can be tried out in a simulation, its effects observed, and revisions made where appropriate. One such task might be to design a set of activities to teach

younger students about Newton's Laws using a *Dynaturtle* (diSessa, 1982; White, 1984). A *Dynaturtle* is moved by firing impulses like a rocket in outer space--controlling an object in a frictionless environment. Students might be evaluated on such a design task in terms of the creativity of the design; understanding of subject matter; systematicity or coherence of the design; how well it carries out its intended purpose.

Advantages. The plan to develop simulations as assessment systems is feasible in much of the current school curriculum, allowing for (a) focus on problem-solving skills in which sequences of strategic action can be captured and recorded, (b) the potential of monitoring the ability to learn in a domain as well as prior knowledge, and (c) adapting to students' prior knowledge which enables measurement of generative rather than recognition knowledge. Many simulations--from which exemplars can be chosen--are also practical as assessment environments for the level of hardware currently available in most schools.

Issues. In order to use them as part of new assessment practices, however, there are several critical research and development tasks that need to be completed. We need to determine what kinds of tasks are best suited to use of simulations for assessment--to experiment with tasks that have different kinds of definitions and constraints. Some possibilities are described above, and further exemplified below (see Scenario 1). Second, we need to determine what kinds of records can and should be collected about a student's work in these computer-based environments. This can be as detailed as key-stroke data, or as high-level as notes they themselves write reflecting on their work. The key question is: what kind of data should be recorded for assessment purposes (likely different for different subjects and simulations), and how should this information be compiled for efficient scoring?

Third, existing simulations need to be "retrofitted" to structure the assessment tasks, and to automatically collect appropriate records of students' work. The kinds of records that validly reflect desired skills, and are efficient to handle, have to be determined through analysis and experiment. They may be different for different subject areas and age levels. Fourth, scoring systems need to be developed and repeatedly tested for use with these kinds of records. In this case, the scoring is done by master assessors who are taught to use the system. While these tasks are challenging, and require time and effort, we believe that there is practical payoff in the relatively near future. Successful models of how such assessments are designed and can work in practical settings can serve to expand this approach to assessment in different domains and developmental levels.

4.2.3. Multimedia systems

Multimedia has long been a feature of educational life. For example, teachers and students in typical elementary school classrooms are surrounded by multitudes of materials--paints, blocks, clay, musical instruments, costumes, animals, plants. They often work in mini-worlds within the classroom, each of which may use different types and compositions of materials, machines, activities, or instructions. In these environments, the unique and powerful features of the different media can be used to their particular advantage and in combination can create the kinds of complex educational experiences now advocated.

Recently, the electronic integration of the different media (video, graphics, text, sound) has made possible new multimedia opportunities for instructional environments (Ambron & Hooper, 1987, and new--as yet relatively unexplored--opportunities for assessment. These developments that allow multiple-media to be stored and orchestrated on a single disk have ushered in an era of considerable experimentation in design and development. Yet the technology is still too expensive for the average classroom, and it is too early for definitive views about its best roles and effects. A system to support the use of the DVI product, *Palenque* (which allows users to explore in surrogate fashion the Mayan archeological site and consult a variety of visual databases, Wilson, 1987, and see below) today costs approximately $20,000. Its most immediate use is in public spaces like museums. We can anticipate that these costs will be reduced in the future.

Assessment and multimedia. As the systems are designed for instructional use, simultaneous attention should be paid to the problems of design as environments that support new assessment practices. As with simulations, the key problems concern task structure, record collection and compilation, and the development of new kinds of scoring procedures. One field experiment in this regard is now beginning as part of the *Jasper* Project at Vanderbilt University. *Jasper* is a multimedia system that enlists students in solving problems through dramatic video segments, and then provides information through multiple linked databases for working on those problems (see Appendix for details). *Jasper* is now being integrated into science and math in a number of schools, and the Project is considering how it can support project-based assessment practices. Additional research and development projects concerning multimedia design and use in schools need to explicitly undertake the problems of assessment as part of the research and design agenda.

Currently, Kathy Spoehr (Brown University, personal communications) is investigating how hypermedia environments can be used for new assessments. She is particularly interested in how students' navigation through linked materials may reveal the sophistication of their knowledge in a domain when compared with the performance of experts.

As environments that support new assessment practices, multimedia is best considered in three ways:

First, these systems can function analogously to simulations, structuring problems or inquiries for students (see above), providing them with information to solve the problems--in this case in a variety of modes and media--and recording the processes they engage in as they work. The advantage is that these problems can be closer to real life situations (through, for example, the creative use of video material), and promote the needed skills of interpretation of multiple information types.

A hypermedia system being designed for teacher education in mathematics (Lampert & Ball, 1990, see Appendix) is also interesting for these types of tasks because of its design. Using the system, students look at samples of video records of real events (in this case, mathematics lessons in different classrooms). These are combined with text windows and graphic overlays to explain or comment on the events. Students can be given problems to solve using the system-- to make a decision about a course of action (e.g., what next steps to take with a student struggling to learn fractions). This design could be adapted for student assessment tasks in a variety of domains to create event-based decision-making tasks.

Multimedia systems can support the collection of new kinds of evidence about students' thinking. For example, *Palenque* and *Jasper* provide a complex body of visual information that students can use to solve problems or conduct inquiries. They must view this information, identifying what is relevant for their purposes and then applying it to particular problems. Students' consultation of the visual databases as they work on a problem can be compiled into a record of their actions. In *Jasper* (Young, Haneghan et al, 1990; Young, Vye et al, 1990) the multimedia database supports mathematical problem-solving, and allows a view of students' selection and application of information (sorting relevant from irrelevant) to solve real-life mathematical problems.

Second, multimedia environments can be used to provide a common generative task for students that can have narrative and real world properties. Given the scarcity of the machines in schools, the systems can be used to structure tasks and provide information sources for the whole class. Students then solve the problems that have been posed to the group using paper and pencil or other resources. The records of students' work can be collected through different media than the costly multimedia systems (either computer tools or pencil and paper).

Third, the current level of development of multimedia systems for education can provide composition tools to students. New approaches to assessment advocate the use of student productions/projects/creative work over time to serve as the basis for evaluation. In addition to pursuing lines of inquiry through the material, in some of the systems students can make use of the available information to compose their own products or productions. This makes different kinds of student products available for assessment purposes, and because students are creating their productions in a closed information space, traces of their composition process in choosing and composing information can be available as records. Thus, for example, in *Palenque*, students choose from the system-resident material, yet a composition task can be open-ended and generative in the selection an organization of this material (e.g. make a "movie" about the interpretation of hieroglyphs at Palenque).

Advantages. Multimedia systems can structure interesting problems for assessment purposes--problems that are likely to be closer to real life than is possible with other media because of the flexible visual capabilities. They also offer expanded information resources to students to use in solving problems. Composition as well as inquiry processes can be traced. In initial tests, these systems pr ve to be quite motivating to students.

Issues. There are relatively few exemplars of multimedia software currently. Most of these are design examples or research prototypes. Many of the multimedia examples are created using HyperCard and videodisk, but these are not fully integrated systems and student use is difficult to track. The hardware for fully integrated systems is also still quite expensive, and unlikely to appear in sufficient numbers in all but the wealthiest schools in the near future. As with other types of software, the systems are being designed with little thought about or conversation with those concerned with new forms of assessment. Problems concerning type of task, record collection, and scoring need to be addressed for this type of technology as with others. It would be to everyone's advantage to consider these issues in this relatively early development phase.

4.2.4. Tool software

Tool software, imported initially from offices and laboratories, has played a central role in the development of software for education (Bloomberg, 1986). A tool need not be cognitive in nature to be used to augment cognitive task performance. Most current word processors, databases, spreadsheets, graphics programs are tools that can enhance and transform cognitive tasks, without incorporating models of users or tasks that would make them "intelligent". Tools provide not just quicker and more efficient means for accomplishing textual or mathematical operations, but also provide dynamic help to learners by embodying "procedural support" that provides structure and hints for complex thinking tasks such as writing (Pea & Kurland, 1987), guiding students through inquiry into progressively more challenging encounters with content material (Hawkins & Pea, 1987), or conducting statistical analyses (Rubin, Rosebery & Bruce, 1988).

The sequence of activities defined by tools helps students to recognize task components and to explore materials and ideas. Effective use of tools needs to be coupled with rich configurations of instructional resources. These include resources currently available in schools, such as texts and classroom experiments, but also includes richer supplies of texts than are commonly available, visual and graphic information, numerical data and resources outside the schools. They need to be used in conjunction with a rich activity and resource base.

In general, the classes of tool software that have potentially important roles to play in assessment include word processors and writing environments, database software, mathematics tool programs (including, for example, spreadsheets and specifically created instructional tools like ELASTIC for statistical reasoning).

Assessment and tools. Tool software has its primary value in being able to record the process of students' work over time in situations that are more diverse and less constrained than simulations or multimedia (where content resources are often system resident, and traces of students processes in working with this information can be recorded in real time). In contrast to real-time traces, "snapshots" of students processes as they record information with the tools in a sequence of work sessions can provide interesting records of their progress. What is being recorded is the evolving product--its development and revision over time. These tasks can be more or less constrained by directions external to the software (e.g. a word processor to record the stages of development of an essay in history, or a spreadsheet program to support the solution of a multi-stage problem in mathematics).

Tools allow for the efficient accumulation of a record of project-development over time, not as a series of traces of students' actions, but in the form of revisionary activity over sessions. This also permits efficient storage of often cumbersome and interleaved material. Because it is generic, tool software tends to be less intrusive in task definition than are simulations--although specifi problem guidelines can certainly be written into generic tool shells.

The tasks are defined externally to the software, yet the properties of the software embody procedures that guide students' actions. One interesting and potentially recordable feature of these interactions is students' ability to carry out specific subprocedures in a complex composition (e.g. in a writing environment, their success in generating topic sentences; in a database project, how they handle and/or constructions, and the like). Because such technology-based environments support accumulation of records over time and revision, they can well-suit portfolio approaches to student assessment.

Advantages. Tool software is widely available, relatively inexpensive, and runs on all levels of hardware. It generally forms the backbone of a school's software collection and the basic tools are now familiar to a large number of teachers and students. Tools can support and extend the kinds of records that can be collected for portfolio or process-folio approaches to assessment--making them sensitive to revisionary aspects of students' work. They may be used to enhance current alternative assessment methods that are most well-developed in the arena of writing assessment. They may thus form the first step in enhancing new assessment methods with technologies.

Issues. Tools in and of themselves to not provide task structures or information sources to students. These must be offered externally. Records and file structures are often cumbersome to handle, especially for novices. Decisions need to be made about what aspects of students' process work can /should be automatically captured for assessment, and how these should be scored. For example, in what ways should revisionary activity be judged?

4.2.5. Video technology uniquely provides the opportunity to collect records of interactive activities as part of the assessment package. Videotape is the key technology in this regard, and emphasizes tasks such as student oral presentations, collaborative work, explanations and querying of each other as potential assessment activities. Each of these kinds of tasks places emphasis on different important aspects of the interactive use and construction of knowledge.

Videotaped presentations, interactions, interviews may be judged for a variety of valued qualities that are not now emphasized in standard assessment practices--the abilities to explain information and present it coherently to audiences, to ask good questions and pursue understanding, to interact cooperatively, to adapt explanations and demonstrations to different learners' needs, to record work at various stages of progress. Video allows social aspects of learning and thinking--now being recognized as key aspects of expertise and performance throughout the domains--to begin to be incorporated into measurement approaches and thus emphasized in schooling. Video as a medium for assessment has the following four advantages.

First, it can record how students explain ideas and answer questions that challenge their understanding. Oral presentation is critical to many aspects of life, and video enables us to emphasize and capture these skills in the same way text captures written presentation. It also enables feedback, reflection, and recurring practice to be part of a measurement scheme. Video allows us to see how well students integrate words and diagrams, how they answer challenging questions from their audiences, how they deal with counter-examples and arguments, how they clarify points that are unclear to the audience.

Second, video can record how well a student listens. It is possible to see how a student listens to other students or adults, how well they ask questions, and critique or summarize what is said. Listening requires a variety of critical skills: communicating that you don't understand, directing the discussion in ways relevant to your needs, elaborating or synthesizing their ideas. It is the only medium that enables evaluation of listening abilities.

Third, video can record how well students cooperate in a joint task. The skills of cooperating and coordinating distributed work on a complex task are critical to almost every aspect of life, and yet they are discouraged in most current school practice. In recording their interactions, it can be used to measure how they work with partners, offer constructive comments, monitor their partners' understanding, coordinate different aspects of a complex task.

Finally, video can record how students carry out tasks and perform experiments. Because video can record students conducting actions, it allows us to evaluate their ability to perform science experiments, use tools, follow instructions, or create new objects.

Interactive activities

Presentations. Students are often asked to present the results of their work to teachers or to other students. This is often the final product of project-based work. Such tasks should be structured to include a presentation and explanation segment, a segment for clarification questions from listeners, and a questioning period when students are asked to defend their arguments and beliefs. Presentations might be judged in terms of clarity; depth of understanding; coherence; responsiveness to questions; monitoring listener's understanding.

Paired explanations. This tasks makes it possible to evaluate students' ability to explain ideas as well as to listen. First one student explains to another about her project or about a concept (e.g. how friction works), using appropriate tools or media. Then the two students reverse roles for another concept. The explainer might be evaluated according to the above criteria, and the listener in terms of quality of questions; ability to summarize information; helpfulness in clarifying the ideas; appropriateness of interruptions.

Collaborative problem solving. Video can be useful as a medium for use in the assessment of students' abilities to work together to solve problems. The tasks can consist of any of the problem-solving situations described above with technologies, science experiments, construction projects, historical projects and so forth. Criteria would be adapted to the particular context, but would include the following characteristics: helpfulness; creativity; understanding; sharing of work; monitoring progress toward the goal.

A central problem that is just beginning to be addressed is how to score these kinds of records in a way that is efficient and manageable. What kinds of criteria can be used to encourage attention to particular kinds of interactive activity, and are also readily and reliably observable from these records?

Advantages. Video allows recording of different kinds of skills and performances than is possible with any other medium. The needed technology is now available in most schools, is inexpensive and easy to operate. In some schools now, students do the recording of each other. The records are transparent to everyone--teachers and students can view them directly, and use them to improve their performances.

Issues. Tasks need to be defined that well-use the potential of video for assessments. Scoring systems also need to be created that are efficient and reliable. To what degree can the performance settings in different classrooms vary, and still sustain reliable scoring? Video records are also cumbersome to handle--both for scoring and storage. Relatively short performance tasks need to be developed for assessment purposes because of the time involved in viewing lengthy performances.

(V) Scenarios: Social and Organizational Aspects of Technology-enhanced Assessment in Schools

Technologies can substantially contribute to realizing new assessment systems and environments. We develop below three scenarios that play out different forms of entwining assessment/instructional environments in schools that take advantage of different of the technologies.

We are all struggling with what it means to create new ways of assessing knowledge development that validly capture skills now of primary value in our educational system, and that reflect current understanding of how learning takes place. Most current efforts in alternative assessment are focused on the central task: what are valid, reliable and practical ways to assess these goals for student achievement and school accountability?

The framework of systemic assessment suggests that there are two other critical issues that must be simultaneously considered: the social organization of testing and instruction; and, deep beliefs about how we as a society understand and represent learning progress.

Social organization of testing and instruction. Our argument is that assessment methods have direct and often unintended consequences on instruction and the social organization of schooling. This is reflected in what is taught and emphasized (e.g fact-drills vs. in-depth projects that emphasize thinking and the application of knowledge to complex problems). But it also has an impact on how schooling is organized.

Currently, formal testing (as opposed to teacher-designed tests as part of the curriculum) is isolated from the daily activities of schooling--both conceptually and practically. Conceptually, because our deeply-held common-sense view of intelligence (see above) is satisfied by the assessment of progress through small decontexted problems in isolated testing settings. Formal tests occupy a different conceptual category than the tests or other indicators that teachers create or adapt to gauge student progress. And practically, testing is generally carried out as a special day or week--quite separate from the ongoing curriculum. Students are commonly specially prepared for this testing occasion, and days are blocked out during the year when daily learning activities are suspended and students participate in special testing sessions.

The new kinds of assessments that are now being explored will likely have a major impact on the social organization of testing. Because they are designed to measure problem-solving and thinking skills in situated ways, these methods tend to require considerable time and effort on the parts of students (and, likely, teachers). They often are designed to take place as extended problem-solving sessions (e.g. some performance-based measures), or as work collected over time (e.g. portfolios). Rather then being isolated and specialized circumstances, there will be considerable pressure to merge the social organization of testing with the social organization of instruction. Testing will likely become embedded in the curriculum, or at least significantly more closely related to the ongoing activity of the classroom. Teachers will have to assume different roles in relation to testing--they may assume more responsibility for creating testing circumstances in the course of daily curriculum activities, and for collecting student materials for external evaluation. But formal tests will become more closely aligned with how the teachers themselves monitor student progress.

These pressures to embed testing in instruction are not yet receiving the attention they will warrant concerning the practical consequences. Will all students end up doing the same complex problems at the same times all over the country? What kinds of variation on task constraints will be allowable in different settings and still allow reliable scoring? If all simulations and multimedia programs come with the ability to collect records about students' progress and suggested tasks, will this promote comfortable variation in assessment circumstances? To what degree can our familiarity with the current social organization of testing--separate from everyday instruction and student/teacher activity---be modified to embed tests in instruction? To what extent can the records of student progress be directly derived from their ongoing learning activities? What will this mean for time and schedule of the school day and year? What will it mean for teachers' roles?

The representation of learning progress. While we will not address this problem extensively, it is important to note another key aspect of the social organization of testing that will likely undergo significant revision. If new forms of assessment are to broadly take hold, public awareness of what educational progress is, how it is assessed, and how it is represented will need to change.

Instead of viewing testing as a special, well-defined cultural occasion/event that has an opaque numerical score or scores as its consequence for representing performance weeks or months later, a transition will be needed to situated testing of complex performances as part of instructional settings. Representations of progress will likely look different from the familiar numbers, at least some taking the form of profiles or qualitative judgments or the like. This has implications for the transparency criterion--that students and teachers know what they will be judged on. Instead of a separate and opaque test, the circumstances and standards of performance are public.

Issues to be considered in the social organization of testing: (1) Situated learning requires students to work on problems that are the same as or related to the context of instruction--not decontextualized performances; (2) These kinds of tests are time-consuming, often individualized to some degree, and can take several sessions; (3) If some of the work is sent to master assessors outside the school, what is the teacher's role in setting up tasks, collecting records, and interpreting student work? How does curriculum complement or merge with the requirements of these new forms of testing? (4) Who are the judges or scorers? How does the educational system accommodate the training and financial requirements of scorers (budgets, unions, locations, etc.)? Will this be the responsibility of private testing organizations?; (5) Public understanding of the purposes, methods, and representations of assessment and the nature of mind needs to be addressed.

The scenarios below present three different views of the use of technology-enhanced assessments in schools, with particular attention to the social and organizational features of schooling. What are some anticipated consequences of using different technologies that can embed, at least partially, testing in instruction? We begin with the least radical--most realistic--scenario for technology-enhanced assessment in the relatively near future. Scenario 1 is based on the technologies that can support performance-based and portfolio tasks and record collection (simulation, video, tools, multimedia) combined with human judgment/scoring. Here, testing is only partially merged with the ongoing work of school. Scenario 2 is more future oriented, and explores the organization of learning and testing with support of intelligent tutoring systems. Scenario 3 is the most radical, and discusses combining all available technology resources to create seamless teaching/testing environments. This requires both considerable design advance for systems that merge instruction and assessment, and a reorganization in how testing is socially organized and understood.

5.1 Scenario 1

This scenario is the most feasible approach to realizing systemically valid assessment on a broad scale in the near term. We believe that this realization requires the use of technologies to reveal aspects of performance that are required for high quality education, and to make them practical in schools. Appropriate assessment design and technology use in the context of this scenario can successfully meet the four criteria of systemically valid tests. The scenario is especially relevant for the directness and scope properties.

The ideal assessment system is designed to cover all knowledge and skills that are valued as goals of schooling (scope criterion). An assessment system that is faithful to the broad scope of learning outcomes known to be critical today must assess a range of other processes and performances than has traditionally been the case--with respect to both domain knowledge and use. The valid assessment of this range relies on considerably more complex and multiple tasks and assessment occasions than the current dominant paper and pencil technologies (e.g. key features of alternative assessments, see above). In this scenario, technologies play the key role in specifying, organizing and recording a variety of aspects of student performances and productions that were previously inaccessible. These records are then made available to procedures for scoring by master assessors and teachers.

Such an approach to assessment requires several components, regardless of the domain. First, a set of tasks must be defined that are representative of the knowledge and skills expected of students. The roles of technologies in structuring these tasks (content and procedures) must be specified, and variations, levels of difficulty, provisions for feedback and practice/revision defined. The tasks can be defined within the technologies themselves (e.g. simulation programs), or they can be defined externally to the systems which are then used to facilitate students' work and/or collect detailed records not otherwise available (e.g. tool software and videotape).

Second, the particular records of student work that are useful for different domains and tasks must be specified (e.g. are sequences of actions collected through trace procedures useful, and at what level of compilation?). Third, libraries of exemplars that include the specified records for a range of students' performances need to be accumulated for scoring development and training purposes. These are also needed for the transparency criterion to make clear to teachers and students the criteria for assessing performances. Technologies can be enlisted to help students to improve their performances. They should practice their performances, making use of coaching or feedback that can be built into the software. They can also examine and critique their own and others' videotaped performances. Students need to practice and receive feedback about these performances, that include consistent and detailed specification of the various benchmarks of high quality performance. What constitutes good performance needs to be clear, detailed, and exemplified.

A wide variety of tasks and performance choices are possible with the approach in this scenario, supporting complex profiles of student accomplishment. Carefully designed compositions of assessment tasks can ensure efficient coverage of large parts of the curriculum for complex learning goals. The design of the various collected records needs to be coordinated with efficient scoring/judging procedures. This scenario lacks the specificity of problem-solving guided by models of expertise embodied in intelligent system-based visions of assessment. But it has the advantage of addressing the scope and directness criteria that are problematic with intelligent tutoring systems.

Many of the kinds of records proposed require subjective scoring, which are often objected to as costly, time consuming and inherently unfair. However, subjective, primary trait, clinical methods are currently argued for by many of those involved in the development of alternative assessments as those most likely to promote the kinds of instructional contexts we seek, and to be the only means available to capture the most important aspects of students' accomplishments. As argued elsewhere (Frederiksen & Collins, 1989; Mullis, 1980), methods of achieving fairness in assessing student writing are well developed, and these methods are readily adaptable to records of thinking processes, complex products, and interactions, supported by computers and video. The limits of what we know how to objectively score may so fundamentally misdirect the educational enterprise that the real costs of objective scoring may outweigh the costs of an assessment system designed to measure a broad range of student abilities that are exhibited in complex tasks.

This scenario involves the role that technologies can play in performance-based assessment (including portfolio and other complex tasks) that are subject to human scoring and judgment. A variety of technologies are used in the process of collecting data about students' performances in the instructional context, and are then submitted to scoring procedures external to the system. The technologies can support three different kinds of assessment activities that provide new means to understand aspects of student performance previously inaccessible with traditional technologies.

These include: problem-solving activity; compositional activities; interactive activities (see simulation, multimedia and video, above). Each of these kinds of activities can be entwined in the instructional context, yet constrained sufficiently to provide consistent data for reliable scoring.

Many of these records will be new to broad uses in assessment systems (e.g. trace data about problem-solving, video records, computer-based process-folios). There are significant issues to be resolved in handling such records, including: what aspects of performance and level of detail is appropriate (e.g. which performance traces have significant meaning--different for an human judge than an ITS)?; how are these records best compiled and annotated?; are there critical performance aspects that can reduce the amount of data?; and the like.

Scoring systems indexed to these kinds of records need to be developed, built on a small number of essential criteria for performance in the tasks that can be readily understood by students and teachers.

This scenario assumes that the curriculum will be organized to include significant time spent on complex or extended problem-solving activities and student project-based work. In the systemic assessment view, this accommodation of different kinds of student activity than is practiced in many classrooms may be driven by new assessment practices. The scenario further assumes that students have regular access to computer resources but this does not demand significantly more equipment than is now available in most schools. Classrooms that have 4 to 6 computers could support these kinds of assessment tasks, or technologies organized into resource rooms in which students carry out their independent work. Use of video-recording is also envisioned as part of assessment practices in this scenario. Sufficient video cameras and recorders are also needed for regular videotaping of performances (i.e. brief taping sessions of students every two to three months)

For example, physics students in a high school are assigned to carry out a complex project over the course of three months. The project enables students to develop and use the basic physical science and mathematical ideas that are the focus of the curriculum. The project absorbs significant time, and is done in parallel with class discussions and problem-solving sessions. The featured project this trimester concerns projectile motion. Each student must select a projectile, and create a set of products that display her understanding of the target concepts in multiple contexts. The science classroom contains a group of six Macintosh computers that students use for their projects. Students can also use the 25 computers that are available in the library. The project is organized so that students use the machines at different times in their project development. Thus, the computer-enhanced testing that is embedded in the project is not carried out simultaneously by all students; use of the scarce resource can be distributed over time.

For their projects, students must create representations in different symbol systems and media that demonstrate or explain projectile motion. They must use a simulation program to create a model of the motion of the projectile and include tests of the effects of a list of key variables (like friction). They must describe the physical motion principles exemplified by their projectile in several symbol systems: graphs, equations, text explanations. They must explain their understanding and present their projects to an examiner, and to another student. Some parts of the student's work is used for diagnosis by the teacher. Other aspects of her work are collected to be externally judged. The "formal" assessment of each student's progress takes key slices of the project work, and compiles these records of students' performance.

In this case, the test consists of records of students' problem-solving processes as they work with the simulation. The test is designed to present a task and to automatically collect records of certain decisions and actions. For example, a student's decisions about how to test and record information about effects of variables are important evidence about mastery of scientific methods. Videotape is also used to record aspects of performance for assessment. Students' explanation and listening abilities are judged through performance on a structured task where they must explain concepts to each other are taped. These interactions are structured and brief--practical as assessment records--yet feature data about students' level of understanding of the phenomena. Students' final products are collected as print records. These three types of records (process data from the simulation, videotapes of explanations, and project products) are collected and submitted to master assessors who score them for key features of understanding.

In another school, eighth grade students are studying genetics as a major unit in their science curriculum. In the first weeks of the unit, they have studied genetics concepts through books, videos and other resources, and teacher guided discussions. They have also conducted an important hands-on experimental project in which they bred fruitflies. At the conclusion of this real experiment, a computer-based program that simulates genetics experimentation with the flies is introduced. This is used as a teaching/testing environment. Students now are asked to try out their ideas and hypotheses about genetics concepts in a simulated environment that enables them to get instantaneous feedback. Problems can be customized to relate to the curriculum they just studied. The simulation automatically collects students' decision-making processes as they work through the assigned problems, as records for assessing their progress. How well have they assimilated the ideas? How do they put the ideas to work on related problems? How do they use the information and feedback provided by the system? These records are compiled and scored by the teacher or by master assessors. The testing grows out of, and also extends, the instructional context since students are continuing to learn as they work with the test items.

This approach to assessment is partially embedded in instruction, but structures the testing around meaningful, situated problems that continue to advance students' learning of the subject matter.

In a sixth grade classroom students are assigned to do inquiry projects and write reports about the Civil War. There are several topics they can choose from; all topics require them to analyze causes of the War. Students work independently on their projects as part of their class time and homework over several weeks. Class time is also devoted to seminar-like discussions of readings and ideas about this historical period. Students use two kinds of computer tools for their projects: a program that helps them to structure their inquiry activities (INQUIRE), and a writing environment (WORDBENCH). The writing environment provides tools for prompting good writing and analyzing it for revision (e.g. help with constructing topic sentences, spell-checker, and so forth). With the inquiry software, students must generate questions about the topic and refine the questions as they gather information. They must create a plan for conducting the project, and monitor it. They must record and categorize information, and systematically interpret it. The writing environment helps them to generate the report.

The software is set up to be a production/testing environment. It will automatically collect records of students' work at particular points--original questions and revisions in them; scheme for categorizing information; their analyses of the information using the software formats. Their writing is also collected in a way that annotates its revisions. These records are compiled as a project portfolio along with the final product and submitted to assessors. Key features of the student's thinking and revisionary work are part of the assessment.

In the same classroom--five years later--six multimedia systems are now installed. Students are given a similar assignment, but instead of generating written reports with writing tools, they are now to create multimedia reports about the Civil Rights movement. The system offers a wide range of textual, graphic and video materials (news footage, speeches, interviews, documentary footage) about this period of American history. It also provides them with tools for assembling, editing, and composing these materials into a technology-based report. The system structures procedural tasks that they complete as they work on the project (as above--developing and revising questions, creating a system for assembling and categorizing material, and so forth).

Students' work with the system is automatically collected as they create these intermediate products. But this system also adds the feature of tracking how students use the system-resident information. Their pathways through the information are interesting data that help to diagnose degree of expertise in a domain. These data, along with the final product are collected and submitted to master assessors.

In each of the above examples, students and teachers know when the curriculum-embedded testing is taking place, and the criteria by which they will be assessed. Results from the master assessors are quickly returned to the students and teachers as the basis for improvement.

While this scenario for broad based use of technology in assessment is feasible in the relatively near future, several issues need to be resolved.

The kinds of software that can embody the key performance objectives need to identified, selected, and tested. In most cases, the software will need to be modified to collect the appropriate records of student work, and to provide nonintelligent feedback or coaching to students about their efforts. The composition of different media and different tasks needs to be carefully orchestrated so that all important aspects of performance are emphasized for scope.

5.2 Scenario 2

This scenario assumes significant change in the social organization of testing and instruction: central responsibility for both resides in intelligent tutoring systems. These instructional/testing environments engage students for much of their learning time and effort. Teachers are managers and coaches of students as they work with the system-posed problems.

This scenario assumes a level of hardware capability that is today available in few schools. Sufficient high-level machines must be available for use by students in classrooms or resource rooms for extended tutorial/testing sessions throughout significant parts of the school year. We anticipate that in the best case, each student has access to a machine at all times. Minimally, we expect that 5 to 10 machines (depending on class size and level of effort) would be needed to make this scenario realistic. Additionally, sufficient ITS software would be available in at least science/math/technology subject areas to cover significant portions of the curriculum. Curricula would be organized to interleave significant portions of time working on ITS with classroom presentations, discussions, and work on problems supported by other materials (computer-based or otherwise).

In this scenario, intelligent tutoring systems are used as intelligent testers in classrooms. They become the devices for administering assessments to students as part of the instruction. An advantage of these systems is that they are well suited to deliver problem-solving tasks, where problems of differing difficulty can be given to students. As with adaptive testing, the tests start with easier problems and depending on how well the student does, subsequent problem will be easier or harder. As they work, students are given cognitive feedback on how best to solve these kinds of problems so that the full capability of the tutoring system to teach can be used as part of the testing procedure. The test can then measure not simply prior ability to perform the kinds of tasks given by the system, but it will also measure how well students learn to perform these tasks given precisely specified feedback and advice. This gives the intelligent testing system the same kind of capabilities for measurement that are called for by the perspective of dynamic assessment.

Intelligent testing systems of this sort require the student to generate sequences of actions that lead to solutions, and the assessment concerns the inferences from these sequences about their current knowledge configuration and its strategic application to varying problems. The systems have the key feature of being able to analyze sequences, and to infer their meaning in relation to specification of different knowledge states. This gives them detailed diagnostic capability, and the potential to assemble and represent complex knowledge about students' level of performance in ways that would otherwise be improbable for teachers or assessors. The systems are, in general, based on a view of human functioning that seeks to precisely specify characteristics of domain knowledge and its strategic use in problem solving. While the tasks are complex, and admit of multiple solutions, they are precisely specified and the goal is to fully account for students' actions.

Scoring in such a scenario can be based on the same kinds of measures now used to evaluate problem-solving. At the basic level, students can now be scored on correctness criteria in solving problems, average time required, number of correct vs. incorrect steps in their attempts, how they recover from errors. But the degree to which the system has a characterization of what expert performances require, it is possible to evaluate students more directly. For example, in the case of the LISP tutor, the system has an idealized problem solving model that consists of 325 production rules, representing its strategies for solving the problems. As the students work on problems, the system can evaluate the degree to which each of these productions is used appropriately. This is a direct measure of how the expert model has been acquired. It might also be possible to determine whether the students abandon productions in the system that represent particular misconceptions.

For the QUEST system, the student's performance can be evaluated in terms of how far along the progression of sophisticated models the student has advanced. The data consists of a list of models within the space of possible models that the student has given evidence of having mastered. The evaluation can also be based on the reasonableness of the steps the student has taken toward problem solution. In adapting QUEST as an intelligent testing system, additional types of useful data collection are also proposed: performance on problems associated with more and less sophisticated models; similarity of problem-solving steps to the current expert model; querying during problem-solving concerning the reasons for taking particular steps; amount of help and effort required to perform well in more sophisticated model environments as a measure of learning ability;

choice of learning strategies supported by the system (e.g. exploratory learning, learning from explanations, inductive), and learning time required in these different modes. These scoring possibilities are suggested as sources of different kinds of evidence about performance that are possible to obtain in the QUEST system, but are not yet operational.

If such intelligent testing systems were available, how would they function in the educational context? Because they are designed to be instructional systems, we imagine that they would require significant amounts of time as part of school work. They can form the core of curricula in the areas for which they are designed. For example, the *Geometry Tutor* was used as a key component of six high school geometry courses in Pittsburgh (Schofield & Evans-Rhodes, 1989; Schofield & Veban, 1986). This field test of the Tutor revealed interesting effects for classroom organization and teacher-student relationships. Focusing learning efforts on interactions with the Tutor had three key effects. The Tutor changed the relative amount of attention given to students of different levels. Teachers spent relatively more time with less able students than they had in more traditional modes, coaching and supporting their work rather than having more advanced students work proofs at the board. It thus has a major impact on the social organization of instruction in the classroom--away from whole-class display of expertise by teachers and advanced students, and toward active engagement in problems by all students.

Second, teachers' behavior shifted toward the role of collaborator rather than that of distant expert. Third, grading practices shifted to emphasize effort on the part of students' work rather than absolute level of achievement. Students' behavior also shifted to evidence more time on task, level of concentration, and self perception that they were working harder.

There is also anecdotal evidence that instruction in these environments can change the kinds of discourse that happens among learners when they are away from the ITS. Adults working with an intelligent tutoring system (in this case, for airplane mechanics) were observed to spend considerable time away from the ITS discussing their work on specific problems they encountered in the system. This kind of discourse that draws on the ITS experience may be an important element in students' advancement (Brown, personal communication). One hypothesized effect of the system is thus that it presents very well specified problems accompanied by a process notation that is communally shared by all learners. They then have a common experience of complex-problem solving that can be explicitly discussed.

In a high school science class, students are struggling to understand electricity. Their work in this science unit is organized by the ITS, QUEST. In this class, there are 10 machines available for 20 students, so each half the class works with the system on alternate days for the two-hour period. Both on and off the system, students' learning is organized by problems that they are asked to solve. In the ITS, they work with a circuit simulator. The system diagnoses their current level of understanding in relation to an interrelated set of mental models for understanding in the domain. It presents problems of the appropriate level and type of difficulty, and offers guidance in the form of explanations or trouble-shooting hints as students work on the problems. Students can also see examples of how experts solved the problem. The system thus offers different kinds of instructional support, based on the needs and preferences of the student. The teacher observes students, and monitors the information that the system collects about their progress. She coaches students as needed, based on her diagnosis of their difficulties.

The system simultaneously functions as an intelligent tester. It diagnoses and students' knowledge level according to the range of models. It determines how quickly students move between levels, and the kinds of support they require to do so. It thus provides information about students' learning strategies as well as their performances. The ITS records students' responses to queries and help it offers throughout the process. It provides immediate feedback to students about how they are doing, and can offer them explanations. The system automatically compiles features of students' performances that can be used by teachers for diagnosis and, as appropriate for accountability requirements.

In a computer science class, students are learning to program in Pascal. For this term's work, the class is using PROUST as the core of the curriculum. Each student is assigned her own machine. The ITS sets programming problems for the student to work on at the appropriate level of difficulty. The job of the student is to write and revise programs that meet the problem specifications. Students thus learn to program by being guided through the construction of programs of increasing complexity. The ITS locates and identifies the bugs in students' programs and explains why these parts of the program are in error. The term's work consists primarily of successfully writing a series of programs for the problems set by the system.

As an intelligent testing system, students are regularly given short program-writing assignments that are tests. The intelligence of the system is then used to score how well students do on these test problems in terms of correct solutions to programming steps. The system can also pose and score another form of problem: asking students to identify bugs in other programs. Students and teachers are given feedback about performance on these test items repeatedly throughout the semester. Teachers use this information to coach individual students.

Thus, in this ITS-based scenario, instruction and testing are almost completely merged. Diagnostic techniques that are built into the system guide instruction and provide data for assessment purposes. Data collection and scoring occur automatically. Teachers' roles shift substantially from primarily instructional to primarily individualized coaching.

A serious drawback of these systems today is that they are by far concentrated in limited regions of mathematics, science, and computer science, as computational techniques provide more leverage in these domains. One key question is thus whether this scenario applies to any other crucial areas of schooling. This scenario requires large amounts of effort and resources to be devoted to creation of intelligent or semi-intelligent systems and it as yet unclear whether that is possible for large areas of learning.

5.3 Scenario 3

The third and most radical scenario with respect to its embeddedness in instructional practice involves the use of technology-based resources to create integrated, seamless learning and testing environments. This, distant, scenario incorporates all types of technology for enhanced assessments. Each type of technology fits a particular niche in the the instructional/testing context. This scenario assumes that the social organization of testing is quite different from what we know today--testing and instruction are completely merged. Teachers and students now routinely make use of immediate feedback about performances. This kind of individualized monitoring of progress replaces the concept of testing--seamless, situated instruction/testing. This scenario thus assumes a significant transformation in our common sense assumptions about learning and testing.

For example as one component of such an environment, the LISP tutor--designed to teach students the programming language--offers an interesting side effect of the teaching. It also collects a record of students' performances that can be analyzed to test various hypotheses about their learning. Each analysis can be a different slice through the data (learning curves, error rates, response times, allowing differences in ability to learn and remember to be factored out). This medium thus enables evaluation to be carried out as a byproduct of students' actual development. Rather than stopping to take a specially arranged test, the testing comes free in the course of learning

In this scenario, students often work with learning technologies individually or in groups, and the teacher's role emphasizes coaching qualities. The scenario assumes a good variety of computational programs, tools, and technology-enhanced resources (e.g. writing coaches, statistical and graphing programs, computer-based laboratories). All the functions of testing might be realized without taking tests per se, avoiding the bad side effects of testing and making learning time better used and more efficient. There are three kinds of measures that occur in this scenario that can be used to carry out the multiple functions of testing.

(a) Diagnosis is distributed between computer. teacher, and students. Some programs, such as the tutors mentioned above, can make local diagnoses, looking for errors at each specific step and giving advice accordingly. Other tutors such as WEST perform much more global analyses of misunderstandings and errors, and still other systems suggest aids to support self-diagnosis by students (Frederiksen & White, 1989). The teacher would be freed up to engage with students much more individually and to build up a better picture of the difficulties that particular students have as they work with the different tools.

(b) Summary statistics about both students' processes and products can be accumulated and analyzed in ways appropriate for different audiences. For example, a report to administrators might summarize how many students completed a particular program or project, how quickly they went about it, and the general patterns of what they learned. A report to parents might describe systems their child worked with, the progress they made, how hard they tried, and any other measures parents may request. Reports to teachers might summarize the kinds of difficulties each student is having, and what progress each made. Data collected on student performance in an environment that is orchestrated to merge learning and testing would allow many different analyses and views of student progress.

(c) Portfolios could be readily created as libraries of each students best work. This idea was likely first used in the Plato math curriculum (Dugdale & Kibbey, 1975) with the Green Globs game used to teach analytic geometry. Students' best attempts to write equations for curves that go through fifteen randomly placed globs on a Cartesian plane are recorded in the computer-based library as performances. This library concept can be extended to personal portfolios in a variety of areas where records are kept of best compositions, or problem solutions. As we move to a society where learning and thinking are critical, basing decisions in part on student portfolios (they may be in part based on the kinds of summary statistics described above) will take into account creativity as well as selectivity.

Moreover, basing decisions on accomplishments rather than simply on measured aptitudes reflects more realistically the way decisions get made in the real world, stressing the production of good work rather than doing well on tests.

This scenario involves moving away from testing per se to analysis of ongoing learning and accumulation of artifacts produced in the course of that learning. Technology-based tools and systems are essential to the feasible operation of such an environment, and different types and combinations of technologies may well fit appropriate niches. The key features of this kind of future scenario for technology-enhanced assessment are less how technologies can be designed to deliver tests, and more how the social organization of testing can become completely merged with teaching/learning activities.

Summary

Significant research, development, and testing needs to be done to create the kinds of assessment methods and practices now needed. We believe that progress in the development and practical realization of alternative assessment methods can be significantly enhanced by taking advantage of the capacities technologies offer. The development of these methods, and the development and integration of educational technologies should be more closely allied than they are today. This may mean creating occasions for conversations. It may also mean making problems of assessment an important component of the development of new software and media.

The nonintelligent technologies are more likely to make broad contributions to assessment in the near future. The development of ITS for these purposes will require significant sums, and may not be possible for significant portions of the curriculum.

The social and organizational issues of testing and instruction also need to be considered when these more complex and interactive skills are the new targets of assessment. It is likely that instruction and assessment will be much more closely merged. This has significance for how curriculum is designed, how testing is organized and scheduled, how student records are compiled, and how teachers' work is organized

A range of issues additional issues arise in discussions about these new approaches to assessment, whether or not they are technology-based or enhanced. The first of these is cost. While these types of tests are more expensive to administer, to the extent that testing and instruction become more overlapping there can be less need for outside-agency tests. Students' actual work in classrooms can be used to monitor their progress--in detail by teachers, and through submissions of portions of the work as records or portfolios to an agency that validates the work through master assessors. The high cost of such testing would likely have the added benefit of less outside-testing being imposed on schools--carefully designed performance-based scenarios or portfolios would replace formal multiple-choice test settings.

Second, if student work is closely monitored as they work with computers, this can have the undesired effect of turning the learning situation into one of surveillance. In addition, if the scoring of process-records is set up so that students are scored according to single correct paths, then this will have the opposite effect of that desired. Rather than encouraging thoughtful multiple approaches to a problem, it would force rigid, rote reasoning. To avoid these undesired effects, tasks and scoring systems need to be designed to encourage multiple and creative approaches to problems, and multiple solutions. In addition, to avoid the problem of surveillance, records that are collected of students' work need to be designed to be meaningful chunks--not every action--and to be used diagnostically to encourage thinking and revision.

Third, the issue of fairness arises. Multiple choice testing technology was in part developed to meet criticisms of subjectivity in judging student performance from many different backgrounds. It was intended to meet standards of objectivity so that all students were judged evenly and fairly. Many debates concerning the cultural bias of items have arisen over the years (culture, gender, language group), as have criticisms that the tests measure a narrow band of skills in formats that favor particular groups of students. Issues of fairness squarely confront the research and development enterprise for new assessment methods. But to the extent the new assessment tasks sample a broader range of skills, the opportunity for students from different groups to display competence is enhanced. And from the outset, it is essential that members of the different social and cultural groups have a voice in the design and revision of these new assessment methods.

A range of issues additional issues arise in discussions about these new approaches to assessment, whether or not they are technology-based or enhanced. The first of these is cost. While these types of tests are more expensive to administer, to the extent that testing and instruction become more overlapping there can be less need for outside-agency tests. Students' actual work in classrooms can be used to monitor their progress--in detail by teachers, and through submissions of portions of the work as records or portfolios to an agency that validates the work through master assessors. The high cost of such testing would likely have the added benefit of less outside-testing being imposed on schools--carefully designed performance-based scenarios or portfolios would replace formal multiple-choice test settings.

Second, if student work is closely monitored as they work with computers, this can have the undesired effect of turning the learning situation into one of surveillance. In addition, if the scoring of process-records is set up so that students are scored according to single correct paths, then this will have the opposite effect of that desired. Rather than encouraging thoughtful multiple approaches to a problem, it would force rigid, rote reasoning. To avoid these undesired effects, tasks and scoring systems need to be designed to encourage multiple and creative approaches to problems, and multiple solutions. In addition, to avoid the problem of surveillance, records that are collected of students' work need to be designed to be meaningful chunks--not every action--and to be used diagnostically to encourage thinking and revision.

Third, the issue of fairness arises. Multiple choice testing technology was in part developed to meet criticisms of subjectivity in judging student performance from many different backgrounds. It was intended to meet standards of objectivity so that all students were judged evenly and fairly. Many debates concerning the cultural bias of items have arisen over the years (culture, gender, language group), as have criticisms that the tests measure a narrow band of skills in formats that favor particular groups of

students. Issues of fairness squarely confront the research and development enterprise for new assessment methods. But to the extent the new assessment tasks sample a broader range of skills, the opportunity for students from different groups to display competence is enhanced. And from the outset, it is essential that members of the different social and cultural groups have a voice in the design and revision of these new assessment methods.

References

Ambron, S., & Hooper, K. (1987). Multimedia in education. Cupertino, CA: Apple Computer.

Anderson, J. (in press). Analysis of student performance with a LISP tutor. In N. Frederiksen, R. Glaser, A Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skills and knowledge acquisition. Hillsdale, NJ: Erlbaum Press.

Anderson, J. R., & Reisser, B. J. (1985). The LISP tutor. Byte, 10, 159-178.

Anderson, J. R., Boyle, C. F., & Resier, B. J. (1985). Intelligent tutoring systems. Science, 228, 456-462.

Baron, J. & Sternberg, R. S. (1986). Teaching thinking skills: Theory and practice. New York: Freeman.

Bashkar, R., & Simon, H. A. (1977). Problem-solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science, 1, 193-215.

Beck, I. L., & Carpenter, P. A. (1986). Cognitive approaches to understanding reading. American Psychologist. 41,1098-1105.

Bennet, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1989). The relationship of constrained free-response to multiple-choice and open-ended items. Educational Testing Service RR-89-33, 1989.

Bennet, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (199). Toward a framework for constructed-response items. Educational Testing Service RR-90-7, 1990.

Bennett, R. E. (1990). Toward intelligent assessment: an integration of constructed response testing, artificial intelligence, and model-based measurement. Educational Testing Service RR-90-5, 1990.

Bereiter, C., & Scardamalia, M. (1986). Schooling and the growth of intentional cognition: helping children take charge of their own minds. In Z. Lamm (ed.), New trends in education. Tel-Aviv: Yachdev United Publishing Co., Hillsdale, NJ: Erlbaum.

Bracey, G.W. (1989). The $150 million redundancy. Phi Delta Kappan, May 1989, 698-702.

Brown, A. L., & Campione, J. C. (1986). Academic intelligence and learning potential. In R. J. Sternberg & D. K. Dettermann (Eds.), What is intelligence? Contemporary viewpoints on its nature and definition. New York: Ablex (Special Monograph)

Brown, A. L., & Day, J. D. (1983). Macrorules for-summarizing texts: The development of expertise. Journal of Verbal Learning and Verbal Behavior, 22, 1-14.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-42.

Campione, J. C. & Brown, A.L. (1985). Dynamic assessment: One approach and some initial data. Center for the Study of Reading, Technical Report No. 361, Urbana-Champaign, Illinois: University of Illinois.

Chi, M., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1987). Self-explanations: How students study and use examples in learning to solve problems. Pittsburgh, PA.

Chipman, S. F., Segal, J. W., & Glaser, R. (Eds.) (1985). Thinking and learning skills: Current research and open questions. (Vol. 2)
Anchored instruction and science education. Cognition and Technology Group at Vanderbilt. In press.

Cognition and Technology Group at Vanderbilt. Anchored instruction and science education. In press.

Collins, A. (1990a). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.) Diagnostic monitoring of skills and knowledge acquisition (pp. 75-87). Hillsdale, NJ: Erlbaum.

Collins, A. (1990b). Cognitive apprenticeship and instructional technology. In L. Idol & B. F. Jones (Eds.) Educational values and cognitive instruction: Implications for reform (pp. 119-136). Hillsdale, NJ: Erlbaum.

Collins, A. (in press). The role of computer technology in restructuring schools. In K. Sheingold & M. Tucker (Eds.), Restructuring for learning with technology.

Collins, A., & Brown, J. S. (1988). The computer as a tool for learning through reflection. In H. Mandl & A. Lesgold (Eds.), Learning Issues for Intelligent Tutoring Systems (pp. 1-18). New York: Springer-Verlag.

Collins, A., Gentner, D., & Rubin, A. (1981). Teaching study strategies (Tech. Rep. No. 4794). Cambridge, MA: BBN Laboratories.

Collins, A., Hawkins, J. & Frederiksen, J. (1990). Technology-based performance assessments. Paper presented at Symposium on

Technology-sensitive performance assessment, American Educational Research Association, Boston, MA.

di Sessa, A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. Cognitive Science, 6, 37-76.

Dugdale, S. & Kibbee, D. (1975). The fractions curriculum. Champaign-Urbana: University of Illinois, Plato Elementary School Mathematics Project.

Feurzeig, W. (1986) Algebra slaves and agents in a Logo-based mathematical array. Instructional Science, 14, 229-254.

Frederiksen, J. & White, B. Y. (1990). Implicit testing within an intelligent tutoring system. To appear in Machine Mediated Learning.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Frederiksen, J.R. & White, B .Y. (1990). Intelligent tutors as intelligent testers. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 1-25). Hillsdale, NJ: Erlbaum.

Frederiksen, N. (1984). The real test bias. American Psychologist, 39(3), 193-202.

Glaser, R. (1984) Education and thinking: The role of knowledge. American Psychologist, 39, 93-104.

Greeno, J. G., & Simon, H. A. (1986). Problem solving and reasoning. In R. C. Atkinson, R. Hernstein, G. Lindzey, and R. D. Luce (eds.), Stevens' Handbook of experimental psychology (revised ed.). New York: John Wiley & Sons.

Haney, W. & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. Phi Delta Kappan, May 1989, 683-687.

Hawkins, J., Collins, A. & Frederiksen, J. (1990). Interactive technologies and the assessment of learning. Paper prepared for the UCLA Conference on Technology Assessment: Estimating the Future, September 1990.

Hawkins, J. & Pea, R. (1987). Tools for bridging the cultures of everyday and scientific thinking. Journal of Research in Science Teaching, 24(4), 291-307.

Johnson, W. L., & Soloway, E. (1985). Proust: An automatic debugger for Pascal programs. Byte, April 1985, 179-190.

Lampert, M. & Ball, D. L. (1990). Using hypermedia technology to support a new pedagogy of teacher education. The National Center for Research on Teacher Education, Issue Paper 90-5.

Littman, D., & Soloway, E. Evaluating ITSs: the cognitive science perspective. To appear in Intelligent Tutoring Systems, L. Erlbaum and Associates Press.

Mullis, I.V.S. (1980). Using the primary trait system for evaluating writing. National Assessment of Educational Progress Report. Denver, CO: Education Commission of the States.

Nickerson, R. S., Perkins, D., & Smith, E. E. (1985). The teaching of thinking Hillsdale, NJ: Erlbaum.

Pea, R. D. & Kurland, D. M. (1987). On the cognitive effects of learning computer programming. In R. D. Pea & K. Sheingold (Eds.), Mirrors of minds. Norwood, NJ: Ablex.

Pea, R. D., & Soloway, E. (1987). Mechanisms for facilitating a vital and dynamic education systems: Fundamental roles for education science and technology. Washington, DC: Office of Technology Assessment.

Pea, R. D., (1987). Human-machine symbiosis: Exploring hypermedia as new cognitive and cultural technologies. Unpublished manuscript.

Resnick, L. B. (1987a). Education and learning to think. Washington, DC: National Academy Press.

Resnick, L. B. (1987b). Learning in school and out. Educational Researcher, 16(9), 13-20.

Rubin, A., Rosebery, A., & Bruce, B. (1988) Final report: Reasoning under uncertainty. (Tech. Report No. 6851). Bolt, Beranek & Newman, Cambridge, Mass.

Scardamalia, M., & Bereiter, C. (1986). Computer-supported intentional learning environments. Unpublished manuscript.

Scardamalia, M., & Bereiter. C. (1984). Development of strategies in text processing (occasion paper No. 3). Ontario: Ontario Institute for Studies in-Education, Centre for Applied Cognitive Science.

Schofield, J. W. & Evans-Rhodes, D. (1989). Computers in the classroom: The impact of computer-usage and teacher-student behavior. Presented at the "Computers in the Classroom" symposium at the meeting of the American Educational Research Association, March 1989.

Schofield, J. W., & Verban, D. (1987). Computer usage in the teaching of mathematics: Issues which need answers. In D. Grouws & T. Cooney (Eds.), The teaching of mathematics: A research agenda. National Council of Teachers of Mathematics. In press.

Schofield, J. W., & Verban, D. (1988). Barriers and incentives to computer usage in teaching. Technical Report #1, Learning Research and Development Center, University of Pittsburgh.

Sebrechts, M. M., Bennett, R. E., & Rock, D. (1990). Machine-scorable complex constructed-response quantitative items: agreement between expert system and human raters' scores. ETS Research Report, in press.

Schoenfeld, A. H. (1985). Mathematical problem solving. (Orlando, FL: Academic Press.

Schoenfeld, A. H. (in press). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In D. N. Perkins, J. Segal, & J. Voss (Eds.), Informal reasoning and education.

Snyder, T. & Plamer, J. (1986) In search of the most amazing thing: Children, education and computers. Reading, MA: Addison Wesley.

Spoehr, K. & Katz, D. (1989) Conceptual structure and the growth of expertise in American history. (Working paper) , Brown University.

Tinker, R. (1987) Network science arrives: National Geographic-coordinated school experiment generates student scientists. Hands On, 10(1), 10-11.

Tuma, D. T., & Reif, F. (Eds.). (1980). Problem-solving and education: Issues in teaching and research. Hillsdale, NJ: Erlbaum Press.

Voss, J. F. (1986). Social Studies. In R. F. Dillon & R. F.Steinberg (Eds.), Cognition and Instruction. New York: Academic Press.

Wenger, E. (1987) Artificial intelligence and tutoring systems. Los Angeles: Morgan Kaufmann.

White, B. Y. (1984). Designing computer activities to help physics students understand Newton's laws of motion. Cognition and Instruction, 1, 69-108.

White, B. & Horowitz, P. (1987) ThinkerTools: Enabling children to understand physical laws. (Tech. Report No. 6470) Bolt Beranek & Newman, Cambridge, Mass.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, May 1989, 703-713.

Wiggins, G. Happily taking and teaching to the (authentic) test: Instructionally-supportive assessment. Manuscript.

Wilson, K. (1987) The *Palenque* optical disc prototype: Design of multimedia experiences for education and entertainment in a nontraditional learning context. (Tech. report No. 14) Center for Children and Technology, New York, N.Y.

Wolf, D., Bixby, J., Glenn, J., Davidson, L., & Gardner, H. (1990). To use their minds well: investigating new forms of student assessment. In press.

Wolf, D.P. (1989). Portfolio Assessment: Sampling student work. Educational Leadership, April 1989, 36-39.

Young, J., Haneghan, J. V., Barron, L., Williams, S., Vye, N. & Bransford, J. (1990). The Jasper series: an experiment with new ways to enhance mathematical thinking. Prepared for the American Association for the Advancement of Science. In press.

Young, M. F., Vye, N. J., Williams, S. M., Haneghan, J. V., Bransford, J. D., & Barron, L. C. (1990). Research on videodisc macrocontexts to enhance problem solving instruction for middle school students. Prepared for a Poster Symposium Session at the Annual Meeting of the American Education Research Association, April 1990.

Zuboff, S. (1988). In the age of the smart machine: The future of work and power. New York: Basic Books.

APPENDIX

4.2.1 Intelligent Tutoring systems

PROUST (Program Understander for Students)
Eliot Soloway and collaborators

Computer Science – Pascal
Finds nonsyntactic bugs in Pascal programs

PROUST is designed to identify conceptual errors of novices in Pascal as they learn to write programs. The program understands problem specifications in terms of the goals that must be achieved for those problems, and uses a knowledge base of plans that students know for achieving those goals. Thus, it can locate in students' code the plans for each of the goals in the problem specifications. It grows out of a research program on programming expertise, and is based on intention-based diagnosis. The program matches inferences about students' intentions to deep-structure goal-plan representations.

Output: PROUST identifies student bugs and generates a list accompanied by a statement of how the bug violates specifications of the assignment. It can also identify what bugs are important for various program parts.

Tasks: Students write and revise Pascal programs to meet program specifications (e.g. rainforest problem). Students can also be given tasks to identify bugs in programs generated by others. In each of these cases, their performances can be automatically evaluated by the system.

Some research has been done to adapt PROUST as an intelligent testing environment. A first step was to determine how well the system did in relation to human scoring in evaluating student performance on well-structured tasks. Students were first given the assessment of writing short programs (like the rainforest problem), and their performances were evaluated by the system and by experts in terms of correct percentage of solutions analyzed. There were moderate to high success rates for the system in this relatively open ended task (70%). A more constrained task was then developed to see if the machine percentage could be improved. In this case, students were asked to correct faulty programs that they were given, and the agreement between automatic and machine scoring reached a high level.

QUEST (Qualitative Understanding of Electrical System Troubleshooting)

Barbara White and John Frederiksen

Qualitative models for basic electricity concepts
Troubleshooting for solving electrical problems
(predicting behavior, designing and modifying circuits, troubleshooting)

QUEST is designed around the view that learning consists of an evolutionary progression of mental models, and that expertise is a interrelated set of linked models representing aspects of domain understanding (qualitative, quantitative, functional, microscopic). It consists of a circuit simulator built around mental models of the domain, a tutor that can offer appropriate explanations, a troubleshooting expert which consists of the strategies and methods for applying domain knowledge in solving problems. The process of conceptual growth involves building on and transforming earlier representations. Learning can be promoted through different processes, such as practice in problem-solving at the current model level or more challenging levels, induction of model properties through this practice, explanations, comparison of ones' performance to that of an expert. QUEST supports these multiple learning strategies such that students can learn in an open ended, exploratory way using the facilities to provide explanation and feedback. Or it can provide instructional sequences, or students can induce models on their own and ask for explanations, and so forth. Tasks can thus be structured in a variety of ways.

QUEST is built around a notion of knowledge representation and acquisition that is not scalar, linear and additive, but concerns a space of possible models and the transformations required to move from less to more sophisticated forms. It is a fundamentally different way of understanding knowledge change than that underlying current assessment practices--ability continuum vs. a range of possible models and their application in problem-solving.

Considerable thought has gone into how QUEST might be used as an intelligent testing system. It is claimed that the system has the potential of assessing knowledge of domain phenomena, problem solving skills, and learning strategies. Designs for assessment of understanding--some of which are now available of some of which remain to be implemented--can be done in three modes: evaluation of performance on current and surrounding models, similarity of steps to current expert model, and querying of students during problem-solving. Records that allow for these analysis might include a list of models in the space of possible models that students have mastered; tracing of problem solving actions, evaluating correctness of result and reasonableness of solution steps in context of the expert problem-solving model; asking students to give reasons for their actions; use of hints supplied by the tutor. In addition, QUEST might be used to analyze how well students learn in relation to different learning strategies. One might construct a task where students must learn through explanations, or induction, and measure how readily they gain information is these ways by examining the highest model attained, the rate of information gain, the amount of coaching needed, and the quality of their reasoning.

<u>LISP Tutor and Geometry Tutor</u>

John Anderson and colleagues

LISP programming, and high school geometry proofs

These two tutors grow out the the ACT theory of knowledge acquisition, and are based on a particular view of instruction for skill acquisition in a problem solving context. Declarative knowledge is immediately converted to useful productions. Students receive loose guidance that leads them to successive approximation of the target skills through individualized guidance. Students receive immediate feedback on errors. The tutors use representations of problem-solving expertise to predict students' next steps by comparing the current step to all possible next steps, which results in an interpretation of students' actions. The tutors are designed to minimize load on working memory in these domains by focusing students' attention on conceptual rather than syntactical or operational details. Students can see records of their previous steps by representational notations on the screen.

Students are guided in the tutors in the writing of LISP programs, or the construction of geometry proofs in well-specified but complex tasks.

For assessment purposes, the tutors can collect records of students' performances that offer different slices through the data. Learning curves, error rates, response times can be used to analyze performance. Sequences of steps that students take can be aggregrated and interpreted in relation to optimal sequences of expert problem-solving. Students' ability to learn can be separated from their ability to remember.

4.2.2 Simulations

Where in the World is Carmen SanDiego?

Where in the world is Carmen SanDiego?
Where in the USA is Carmen SanDiego?
Where in Europe is Carmen SanDiego?

Where in Time is Carmen SanDiego?

Subject matter: Geography
History

Grade level: 6 and up

Task: Geography game

Kids, assuming the roles of detectives, pursue criminals, through different geographic locations to their hide-outs;

Kids are given seven days to solve the mystery, the more efficiently they use the clues, the less time do they need to solve the mystery;

Kids need geographic knowledge to make efficient use of the clues--they need to know the currency, flags, animals, artifacts, languages spoken etc. of different countries in order to determine from a list of 3 or 4 locations where to go next;

Kids can used an almanac to help them interpret the clues; the manual gives names and descriptions of the criminals

Kids need to identify the criminal, and trace their travel route to their hide-out in order to successfully complete the game;

At each location, kids can get three clues regarding the criminal's identity or the location to which they traveled;

Records:

1) Number of "hours" to complete a case; number of cases solved per period
2) students' notes about clues and places

3) video records/observations of students using the game

Scoring:

1) read from computer screen; number of cases solved will be saved on disk
   - allows teacher to compare performance differences among kids
   - reflects geography knowledge and problem-solving skills

2) scoring guideline would need to be developed; interesting questions:
   - How do kids make use of the clues? Do they engage in deductive reasoning?
   - Do kids understand the clues? Which clues are difficult and which are easy? (e.g. do kids know about the flags of different countries? teachers could use this info to improve their teaching, i.e. teach flags)

3) scoring guideline would need to be developed; interesting questions:
   - see 2) above
   - How do kids keep track of the clues?
   - Do kids come up with hypotheses, predictions?
   - How do kids make use of resources (i.e. almanac)?
   - If kids work in groups, how do they solve problems collaboratively? Do they dominate? Do they make their ideas explicit? Turn taking in playing the game/entering instructions into the computer?

Instructional Context:

- have kids work in groups to solve problems collaboratively (in conjunction with video taping, this would allow to get verbal records of kids' reasoning processes)

- the game could give students the opportunity to practice collaborative problem solving.

- videos could be used as a basis for discussion of how to do group work.

Physics Explorer

One body and two body dynamic motion
Harmonic motion
Gravity
Waves

Subject matter: Science and mathematics

Grade level: 8-12

Tasks: The Physics Explorer (formerly, in prototype, STEP) allows students to conduct and observe a series of experiments which simulate the behavior of objects and phenomena under different conditions. The student can set and manipulate a set of parameters that control the object and its experimental environment. For instance a student can compare the upward acceleration of an object under different conditions of gravity.

Records: (1) On-screen: experiments that are conducted

(2) Print outs: spreadsheet recording of parameters of physical variables, graphs, text explanations by student in form of note cards, experimental parameters and sequences of decisions

(3) Video records: observations of students interacting with software; explanations of their work.

Scoring

-understanding of interactions among parameters
-appropriateness of experiments conducted
-systematicity of testing of variables
-use of different information sources
-predictions and hypotheses
-interpretation of experiments
-appropriateness of parameters
-group collaboration

The Other Side

Subject matter: Social Studies (American History, World History, Economics), usable also for science (effects of having limited natural resources), language arts (communication skills), psychology (conflict resolution), and current events.

Grades: 5 - 12

Task: The Other Side is a computer simulation game about global conflict resolution. The players in this game assume leadership roles of different nations. The game simulates the relations between two countries in a complex world of limited resources. The goal of the game is to build a bridge between the two countries. To accomplish this goal, the players have to raise money, find different natural resources to make money, and establish relationships with the other country.

The game is suitable for being played in groups of 2 to 6 students each. Each student in a team can be assigned roles, e.g. keyboard operator, map manager, resource manager, yearly planner

The Other Side can be played in different modes:

a) one or two computer version; in the one computer version teams take turns on the computer (the team that is waiting for its turn could use the time for planning); in the two computers version each team of players works on their own computer, ideally in separate rooms;

b) competitive versus collaborative mode; the goal in the competitive mode is to be the first country to complete the bridge; the goal in the collaborative mode is for the two countries to complete the bridge jointly in as few years as possible

c) three different levels of difficulty; the levels require increasing skills to manage economic stability and military responsibility, and differ in a variety of ways such as amount of time players have to execute their moves, the cost of possessions, the amount that can be earned from the natural resources, and the level of acceptable contamination in the world.

Records:

1) Hotline history - displays all the messages which were sent during the game
2) handwritten records (examples attached): map/gameboard, yearly planner sheets, resource manager sheet, economic analysis, notes on the events on both sides (could be taken from different perspectives, such as Journalist, Historian), newspaper
3) video records/observations of students' discussions (gives a record of the process of learning how to collaborate (within a group of players); to trace changes in thinking about conflict solution etc.)

Scoring: Scoring guidelines have to be developed. For each type of records, scoring could focus on:

1) Communication skills

2) Planning, management skills, critical thinking, understanding of multiple perspectives, map reading skills

3) see 2), group collaboration, group decision-making, conflict resolution

SimCity

Subject matter: Social Studies, Environmental Science

Grade level: 6 - 12

Task: SimCity is a simulation that allows the user to plan build, and manage a city. The player assumes the role of the mayor of a city. S/he can build a new city on virgin territory, or by taking control of an existing city (e.g. San Francisco in 1906, just before the great quake). The player's task is to zone land (residential, commercial, industrial), to build the infrastructure of the city (roadways, transit lines, power lines, power plants, seaports, airports, fire and police departments, parks, stadiums), to balance budgets (tax rate, allocation of funds to the fire, police, and transportation departments; each city building operation costs a certain amount of money, e.g. $10 to build a section of a road, $100 to zone one plot of land as commercial), to manipulate economic markets, control crime, traffic and pollution, and overcome natural disasters (e.g. floods, fires, tornados). The software simulates the responses of the citizens in the city; the citizens build houses, condos, churches, stores and factories, and they complain about things like taxes, the mayor, city planners; through messages on screen the citizens inform the player of their need for more housing, better transportation, pollution, parks etc. If the citizens get too unhappy, they move out, and the mayor collects less taxes.

Records:

1) on-screen (examples attached):
   - budget window (tax rate, allocation of funds, cash flow, current funds)
   - graphs window, gives time-based graphs of various city data (residential population, commercial population, industrial population, crime rate, cash flow, pollution)
   - evaluation window, displays public opinion about how good a job the mayor is doing, indicates the worst problems in the city, gives population statistics, and an overall city score which can range from 0 to 1000 (this score is a composite index of many factors;

higher scores indicate a more efficient and successful city)

2) print-out:
   - a city map can be printed out

3) hand written notes
   - students could be asked to keep written records of their budgets, and plans for the city; they could keep a journal to reflect on what they build, the events that happened how they had to revise their plans etc.

4) video records/observations of students working in groups

Scoring:

1) on-screen records reflect to some extent how well the player dealt with and coordinated/interrelated variables [when I worked with SimCity, using an existing city, I was so confused by the map that for the duration of the game I hardly did anything; however, the scores I got were very high – so I don't think that these scores provide much information about what went on in the game]; some of the records indicate which variables were paid attention to, and which were neglected or dealt with poorly

2) a scoring guideline needs to be developed for evaluating the map; possible questions:
   - what would be the quality of life in this city (how close are residential and industrial zones to each other, are there enough commercial zones, utilities, and recreational facilities, is there enough/too much transportation)?
   - what would the economic picture of this city look like (enough industry, infrastructure?)
   - is the design/layout of the city aesthetically appealing?
   - is the terrain used efficiently?

3) again, we need a scoring guideline; possible items for assessment: critical thinking, revision skills, planning

4) need scoring guideline for group collaboration (e.g. problem solving, decision making, planning, coordination)

Notes: Required skills:
- map making and reading skills
- coordination of variables
- allocation of resources
- balancing a budget

- responsiveness to and understanding of the needs of citizens
- planning
- revision
- responses to unexpected events
- decision making
- problem solving

On-screen records and the map do not tend to reflect these skills separately; if teachers are interested in the assessment of any of these sub-skills, it would be best to have students work in groups and to videotape them; video tapes allow teachers to focus on the particular part of the city building process in which students apply the skill to be assessed; in addition, group discussions give teachers access to students' thought processes

Jasper

Subject matter: Math
Grade level: 6 - 9

Task: Jasper is a video disk that contains a series of stories/episodes, which are designed to provide a problem solving contexts {the designers call this a macro context] for students to solve applied math problems. Each story contains an explicit, real world math problem. In order to solve the overall problem, students have to generate and solve several subproblems. The Jasper stories are essentially complex word problems on video. For instance, one story centers around math problems involving the concepts of distance/rate/time. The central character, Jasper Woodburry, is buying a new boat, but the boat has no lights. Since it is already 2:30 pm in the afternoon, Jasper has to make a decision if he has enough time to sail home before it gets dark. The video contains several clues as to how much time the trip takes, how much money he has to buy gasoline, weather conditions, etc. In other words, all the information needed to solve the problem and its subparts is embedded in the story. The problems embedded in the Jasper stories are complex; they require more than 15 steps for solution.

Students typically view the story first as a whole. The overall problem is presented at the end. Students can then use the video disk to search for information needed to solve the problem/subproblems, and which they have previously experienced incidentally. The challenge in their video disk search is to sort out the relevant from the irrelevant information.

Records:

1) potentially students could produce a computer-based report, including video segments that include important information, statements of subproblems, calculations, solution

2) video/observation of students working in groups

3) written calculations and problem solutions

Scoring:

Scoring guidelines, which need to be developed could focus on the following questions:

- Do students plan? How?
- Do students define subproblems? Which?
- Do students attempt to solve the subproblem?
- Do students sort out relevant from irrelevant information?
- How efficient are students in their search for information?
- Do students use the correct math operations?
- Do students solve the problem/subproblems correctly?
- Questions concerning collaboration and problem solving in groups

Instructional context:

- teacher could present students with segments of the story, and ask them to identify which segments contain relevant info and which do not

Palenque

Subject matter: Social Studies

Grade level: 4th to 8th

Task: The Palenque prototype is a multimedia database environment (text, graphics, audio, motion video, still video) about an ancient Maya site in Mexico. The prototype consists of several interrelated components: video overviews (to give users an introduction to the prototype), surrogate travel (to explore the site), a museum database (allows user to browse through information about rainforests, Maya hieroglyphs, Palenque history, maps and aerial views), characters as experts and guides, simulated tools (camera, album, tape recorder, compass, magic flashlight; allow users to collect and save information of interest for later reference), and game-like activities (to encourage the users' interaction with the information). Users can interact with the prototype in two basic ways: a) exploration of information in the database, b) production of a report/presentation using the information in the database.

Records:

1) on screen records: album of pictures taken (pictures can be labelled). Since Palenque is not a finished product, the list of on-screen/technology based records should not be thought of as exhaustive; provisions for obtaining other records could be incorporated (e.g. multimedia records including info from museum, text, video, audio; record of exploration)

2) external records: maps, drawings, block building to reconstruct scenes;

3) video tape/observation of interactions with the prototype of individual kids or groups

Scoring:

1) exploration
   - information seeking
   - decision making

- how do kids make use of resources? do they use resources from all modalities (visual, audio, text?)
- collaboration in groups
- map making and reading
- question asking
- strategies of navigation

2) multimedia productions
- organization
- perspective taking
- content
- group collaboration

instructional context:
- exploration of the database in different modes: as explorer (discovery based), as treasure hunter, as movie maker

Multimedia Databases with Composition Tools

e.g.
MediaWorks
Interactive NOVA
Visual Almanac
Beethoven

Subject matter: Science, Social Studies, Art, History, Music

Grade level: undetermined

Tasks: These programs allow users to browse through multimedia databases of visual, textual, and auditory information, and use composition tools to reorder the information, add their own text and graphics, and create multimedia reports and/or presentations.

Records:

1) on-screen/technology based
   - explorations
   - reports/presentation

2) video/observation of users' interaction (individually or in groups)

Scoring:

1) exploration
   - information seeking
   - decision making
   - how do kids make use of resources? do they use resources from all modalities (visual, audio, text?)
   - collaboration in groups
   - question asking
   - strategies of navigation
   - content related issues

2) multimedia productions
   - organization
   - perspective taking
   - content
   - group collaboration

- handling of/awareness of source materials and their limitations/awareness of the media

Hypermedia environment for teachers in mathematics

Currently, there is a project under development at Michigan State University (Lampert & Ball) that is designed to enlist hypermedia in the education of teachers. Despite the fact that this is intended for adult learners, we will include a brief description of its design here because it offers interesting features for considering the design of multimedia and assessment environments. The system is intended to provide an instructional experience which involves learners in problem-interpretation and solving in the context of real-life performances as mathematics teachers in classrooms. Traditional teacher candidates learn pedagogy, curriculum and domain knowledge and method abstractly, away from the context of real performance in classrooms. The knowledge is thus learned in a way unintegrated with the context of practice, and can remain inert. Candidates then have an apprentice experience in classrooms--often limited to a single master teacher--in ways that seldom allow them to deeply reflect about their class experience and reflectively critique and apply it to multiple types of detailed problems.

The multimedia system is designed to allow students to have experiences that are close to real-life applications of knowledge-in-context. Videotapes of lessons (for example, a sequence of mathematics lessons across several weeks) are assembled and stored on videodisk. The lessons are annotated in a variety of ways through Hypercard stacks that allow students to readily find related episodes, and to see multiple perspectives or interpretations of students' and teachers' performances. For example, students can look at all attempts to deal with concepts of number systems in different lessons, can follow one learner through a series of lessons to track changes in thinking, or can follow a teaching strategy and see how it is used differently and has different results in different contexts. The design also include graphic overlays that present alternative representations of the mathematical ideas under discussion and the like.

This design of the hypermedia system is an interesting model for the construction of instruction/assessment environments in general. It has the interesting feature of allowing students to apply their knowledge to a series of related incidents over time, to record their own reflections through annotations of the video, and to deal with multiple representations/interpretations of presented information. Thus, tasks can be structured that reveal a different kind of knowledge application to real settings, different kinds of records can be collected about a constrained yet rich set of domain events through the annotation ability.